DNA-PKcs/ Artemis Complex in DNA Double-Strand-Break Repair:

Cryo-EM and Biochemical Studies

Shikang Liang

Department of Biochemistry

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Girton College

September 2019

Declaration

This dissertation is submitted for the degree of Doctor of Philosophy and presents a summary of my research carried out at the Department of Biochemistry, University of Cambridge, between September 2015 and August 2019. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except as specified in the text and Acknowledgments. The contents of this dissertation are original, except where specific reference is made to the work of others, and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation does not exceed the prescribed word limit of 60,000 words.

Shikang Liang

September 2019

‘‘If we knew what it was we were doing, it would not be called research, would it?’’

Albert Einstein

1879 – 1955

Acknowledgements

First, I would like to express my great gratitude to my dear supervisor, **Professor Sir Tom Blundell**, for his generous and continuous support on my research project and career plan and for his encouragement, patience and immense knowledge. I am very lucky to have an inspiring mentor like him. Despite his busy schedule, Tom has been approachable and made time for all the group members. He provides advice on not only the research project but also other aspects of being a scientist. His guidance assisted me in all the time of my research and writing of this thesis. In addition, his constant commitment to equality and diversity has been inspiring and strengthened my understanding of their importance. I could not have imagined having a better supervisor.

In addition to my supervisor, many lovely people have provided me with advice and help during the past four years. I would like to very much thank **Dr Takashi Ochi** and **Dr Qian Wu** for their daily supervision during the first two years of my PhD. Their insightful comments and advice on my project promoted my understanding of DNA repair and encouraged me to widen my research and skills. I am also truly grateful to **Dr Dima Chirgadze** for his precious advice on the DNA-repair project and for training me on cryo-EM sample preparation and microscope operation using the Talos Arctica and Titan Krios. The cryo-EM facility team of Biochemistry Department including **Dr Steven Hardwick** and **Lee Cooper** has also been supportive and helpful. Moreover, my collaborator **Dr Taiana Oliveira** from Astra Zeneca and **Dr Dijun Du** helped me with many cryo-EM questions when I first started learning the technique. I have had many fruitful discussions, which I really appreciate, with **Professor Ben Luisi** and **Tom Dendooven**, who have also provided me with many new ideas for cryo-EM experiments. With all their help, my learning of the technique and progress of the project accelerated a lot. Many people assisted me in my drug-discovery project and I would like to acknowledge **Dr Vitor Mendes, Dr Michal Blaszczyk, Dr Sherine Thomas**, **Dr Sheikh Arif** and **Pooja Gupta.** I also appreciate a lot the fruitful collaboration with **Professor Terence Strick**, enabling us to study the temporal organisation of NHEJ in a systematic way. In addition, I would like to express my sincere thanks to other great people who provided me with assistance on learning new techniques during my PhD, especially **Dr Katherine Stott** and **Dr Wei-Guang Seetoh**. I express my gratitude to my graduate thesis panel including **Professor**

Steve Jackson, **Professor Luca Pellegrini** and **Professor Meindert Lamers** for their valuable advice and encouragement. Furthermore, it would be difficult to imagine dealing with all the IT issues without **Graham Eliff**. I am truthfully thankful to each and every **Blundell and Luisi group member** for the friendly environment and all their support.

I would also like to thank all my friends who have enriched my PhD life. My colleague PhD candidates of the same year, including **Johannes Lauenstein, Vincentius Aji Jatikusumo** and **Laura Shen,** have been sharing this journey with me over the past four years. I have also had a great college life with the accompany of the amazing Girtonians including **Huiru Lian, Qi Li** and **Richard Clements**. Moreover, many long-lasting friends supported and encouraged me during the four years of my PhD including **Yan Chen, Ke Liu, Yuchi Zhang** and **Sifan Ouyang**. They made the past four years more colourful.

Last, I would like to express my sincere acknowledgement to my dearest parents, **Professor Shuquan Liang** and **Jianan Kang**, for their enlightenment and unconditional love and support. Without them, nothing would have been possible.

Abstract

DNA is the main carrier of inheritance and DNA damage will thus have serious consequences. DNA double-strand breaks (DSB) are amongst the most lethal forms of DNA damage. One DSB can lead to catastrophic consequences including cancer and cell death. To fix it, there are two main mechanisms-- homologous recombination (HR) and non-homologous end joining (NHEJ).

My PhD project focuses on NHEJ. Unlike HR, NHEJ is not limited by the cell cycle as it does not require a template for recombination. Also, NHEJ has been shown to be the preferred DSB repair pathway in higher eukaryotic organisms including human. NHEJ is dynamic and flexible but can be separated into three steps - DNA end recognition, end synapsis and processing, and end ligation. The main objective of my project is to understand the interaction in the complex between DNA-PKcs and Artemis, which is the major nuclease in the step of end synapsis and processing to help clean up the modified DSB ends caused by external factors including ionizing radiation. It is also the only discovered human endonuclease cleaving hairpin DNA, which is indispensable in V(D)J recombination— a mechanism that provides the immunodiversity of antibodies and T-cell receptors.

During my PhD, I purified Artemis and DNA-PKcs, conduct biochemical and biophysical characterisation of these proteins and used cryo-electron microscopy (cryo-EM) as the main structural method. The nuclease assays for investigation of endonuclease activity targeting hairpin DNA revealed that XLF and XLF/XRCC4 have a stimulating effect on the endonuclease complex without and with Ku. Moreover, I have identified the region of Artemis interacting with DNA-PKcs and collected cryo-EM data for complexes of DNA-PKcs with different Artemis constructs, revealing the interaction mode between DNA-PKcs and Artemis. In addition, I further explored the cryo-EM structure of the DNA-PKcs to provide a firmer ground for modelling and the related complex study.

A part of my PhD project focuses on the preliminary drug discovery of Artemis/ DNA Ligase IV complex. An intrinsically disordered Artemis C-terminal peptide interacts with DNA Ligase IV through concerted folding, showing a site that can be most easily targeted by small molecules. Therefore, during my PhD, different constructs of DNA Ligase IV have been screened and the

fragment-based drug discovery approach was initiated to provide chemical tools or candidate drug molecules to inhibit the Artemis-DNA Ligase IV binding site. Moreover, collaboration work with the Strick group enabled us to monitor the temporal organisation of NHEJ using the single-molecule method, which identified the function of PAXX as an early participating component in end synapsis.

Table of Content

List of figures

Abbreviations

Chapter 1. Introduction

1.1 DNA, DNA Damage and DNA Double-Strand Break

Deoxyribonucleic acid (DNA) is the main carrier of inheritance. It was first isolated and identified in the 1860s by Friedrich Miescher, followed by works of other scientists including Phoebus Levene and Erwin Chargaff to reveal more Chemistry details. Based on the previous Chemistry understanding and the X-ray diffraction pattern of DNA from Maurice Wilkins and Rosalind Franklin, in 1953 James Watson and Francis Crick proposed the structure of DNA. As the substance of inheritance, DNA encodes genes, which are the basic functional and physical units of heredity (https://ghr.nlm.nih.gov/primer/basics/gene). It was estimated by the Human Genome Project that there were around 20000 genes coding proteins but the number has subsequently decreased to 19000. Despite of the complicated interpretation and connection between the molecular basis of inheritance (DNA) and the functional unit of inheritance (gene), there was a clear cycle and flow of genetic information from DNA to RNA to protein. This was first proposed by Francis Crick in 1957 as the "central dogma" of molecular biology. As the starting point and the heart of the central dogma, DNA plays a fundamental role in maintaining normal phenotype and passing on genetic information to new cells. Cells have developed a series of mechanisms to avoid DNA from being mistakenly interrupted as DNA damages or misrepair can easily lead to genome instability due to interrupted genetic information, causing abnormal pathological phenotype.

However dangerous they are, DNA damages happen all the time caused by external factors around us and internal factors within us. External factors include UV radiation and chemical reagents, tobacco smoke for example, and internal factors include replication-fork collapse, reactive oxygen species and self-controlled DNA editing and recombination, V(D)J recombination for example. It was estimated that one cell could experience over 1,000,000 DNA changes per day (Lodish, 2000).

In general, there are five types of DNA damage - base loss or modification, mismatch/nucleotide insertion and deletion, crosslink, DNA single-strand break and DNA

double-strand break. Base loss and modification are the most frequent DNA damage in cells mainly caused by external UV light. Mismatch/nucleotide insertions and deletions are mainly caused by errors during DNA replication process. Crosslinks are damages of DNA caused by reagents reacting with two nucleotides of DNA on the same strand (intra-strand crosslink) or between the two strands (inter-strand crosslink). As the name suggests, DNA single-strand break (SSB) has one strand of phosphate backbone damaged and can be caused by topoisomerase, while DNA double-strand break (DSB) involves breaks in the phosphate backbones of both complementary DNA strands.

Among all the types of DNA damage, DSBs are the most lethal. Unrepaired or misrepaired DSBs will cause significant loss of genetic information and disruption of genome stability, resulting in a series of consequences including cell death and carcinogenesis. All DSBs can be divided to two types, pathological DSB and physiological DSB, and there are various sources for different types of DSB.

In the case of pathological DSBs, the DSB is mainly developed from SSB. For example, ionizing radiation is a main exogenous factor for DSB. It stimulates the production of radiolysis radicals in the cell that would further attack the phosphate backbone of DNA and lead to a SSB (Ward, 1994; Thompson, 2012). When the dose of the ionizing radiation is high, it is likely that two SSBs could happen on the two complementary DNA strands. When those SSBs are close enough, for example in the range of one helical turn, they will result in DSBs (Milligan *et al.*, 1995). To have one such DSB, roughly 10 SSBs are produced by ionizing radiation (Ma et al. 2012). Moreover, pathological DSBs could be developed from SSBs in normal cellular processes. For example, a nick in one strand of DNA could easily develop into a DSB during DNA replication. Actually, the majority of the spontaneous DSBs are developed during DNA replication (Syeda, Hawkins and McGlynn, 2014).

Unlike pathological DSBs, physiological DSBs are mainly spontaneous DSBs cleaved on purpose by the cell. Those programmed DSBs are designed for self-controlled genome editing to increase diversity. A good example is our immune system, which produces the dangerous DSBs and makes full use of it to increase the variety of immune responses. Two important processes of our immune system under this scenario are the V(D)J recombination and class

switch recombination (CSR), creating diversity of antibodies and T cell receptors. In addition to the immune system, DSBs are also induced for recombination of homologous chromosomes from different parental origins during the sexual reproduction (Lam and Keeney, 2015). During this recombination, hundreds of DSBs are made by a meiosis-specific and topoisomerase-II-like endonuclease, Spo11, at non-random hotspots (Keeney, 2008). Moreover, recent research indicated that programmed DSBs also exist in stimulated mouse neuron (Suberbielle *et al.*, 2013).

1.2 Human DSB Response

Although DSBs are not rare and can lead to catastrophic consequences in our cells, there is a highly regulated and delicate system that efficiently senses, controls and repairs all DSBs. The system is known as the DSB response.

Generally, the DSB response is complicated and covers many physiological processes, but can be separated into three steps including DSB sensing, signal transducing and effecting, as previously proposed (Jackson, 2002). Cells first respond to DSBs by detecting them through a series of DSB binding proteins (sensors). The sensors then recruit and activate a list of transducers, which are kinases that amplify the signal through the protein kinase cascade and diversify the original signal of one DSB to many downstream proteins, known as effectors, that moderate all kinds of activities including apoptosis, cell cycle control, transcriptional regulation, increased dNTPs level and the DSB repair.

Three main sensor-and-transducer pathways are involved in response to DNA DSB, including the RPA- ATRIP/ATR pathway, the MRN- ATM pathway and the Ku70/80- DNA-PKcs pathway (Blackford and Jackson, 2017). All the three kinases (ATR, ATM and DNA-PKcs) belong to the family of phosphoinositide 3-kinase (PI3K)-related kinases (PIKKs) and share similar structural features. For example, the kinase domains of all three members are located at the C-terminal region with an upstream FRAP-ATM-TRRAP (FAT) domain and a downstream FAT C-terminal (FATC) domain (Bosotti, Isacchi and Sonnhammer, 2000; Mordes *et al.*, 2008). Also, their whole N-terminal regions before the FAT domain are composed of helical solenoid HEAT repeat domains at different lengths, which regulate the protein-protein interactions of PIKKs with other DSB-response components (Perry and Kleckner, 2003).

In the case of the MRN- ATM pathway, the MRN (MRE11-RAD50-NBS1) complex first binds to the DNA ends at DSBs. ATM is then recruited to the DSB site via the C terminus of NBS1. The mechanism of the activation of ATM kinase activity remains unclear. It was proposed that the MRN complex recruits and activates ATM at the DSB sites as the MRN complex could activate the kinase activity of ATM *in vitro* (Lee and Paull, 2004, 2005). However, there are also cases

of MRN/DSB-independent activation of ATM kinase activity under the circumstances of oxidative stress or chromatin changes (Guo *et al.*, 2010; Olcina *et al.*, 2013).

Although the mechanism of ATM stimulation remains uncertain, it is certain that ATM phosphorylates various kinds of proteins, at the level of hundreds, after stimulation (Matsuoka *et al.*, 2007). It is not understood how many substrates are functionally important but some downstream pathways are proven to be significant for the ATM-induced DDR (Maréchal and Zou, 2013). For example, to amplify the signal, ATM initiates the kinase cascade via phosphorylating other protein kinases including CHK2 kinase. CHK2 kinase could further phosphorylate cdc25, which is a phosphatase that becomes downregulated when phosphorylated by CHK2 kinase. The down regulation of Cdc25 will keep the Cdk (cyclin-dependent kinase) at the phosphorylated deactivated status, interfering with progress through cell cycle. The activated CHK2 could phosphorylate the tumour suppressor p53, leading to induction of p21. The upregulation of p21 will further inhibit the activity of Cdk and slow down the cell cycle.

Unlike ATM that targets at the double-strand sites, ATR sits on the single-strand sites of DSB and is the DDR kinase involved in the DNA replication stress response. ATR is recruited by its partner protein ATRIP, which comes to the single-strand DSB sites via the interaction with replication protein A (RPA) after RPA binds to single-strand DNA directly to form the RPA-coated ssDNA (Zou and Elledge, 2003). After recruitment to DSB sites, ATR kinase activity is further stimulated by other activator proteins, including TopBP1 and ETAA1.

Like ATM, activated ATR also phosphorylates a series of proteins, many of which are also substrates of ATM, indicating that while the upstream situations of DSBs may vary, the downstream pathways may finally converge to have similar responses. A good example is the phosphorylation of checkpoint kinase and the effect of slowing or stopping cell-cycle progression. ATR first phosphorylates and stimulates CHK1 (Hekmat-Nejad *et al.*, 2000; Guo *et al.*, 2000; Liu *et al.*, 2000; Zhao and Piwnica-Worms, 2001). The activated CHK1 then phosphorylates Cdc25A, which will result in its proteasomal degradation. Without Cdc25A, the CDKs will remain the phosphorylated inactive state, which will slow down the cell cycle.

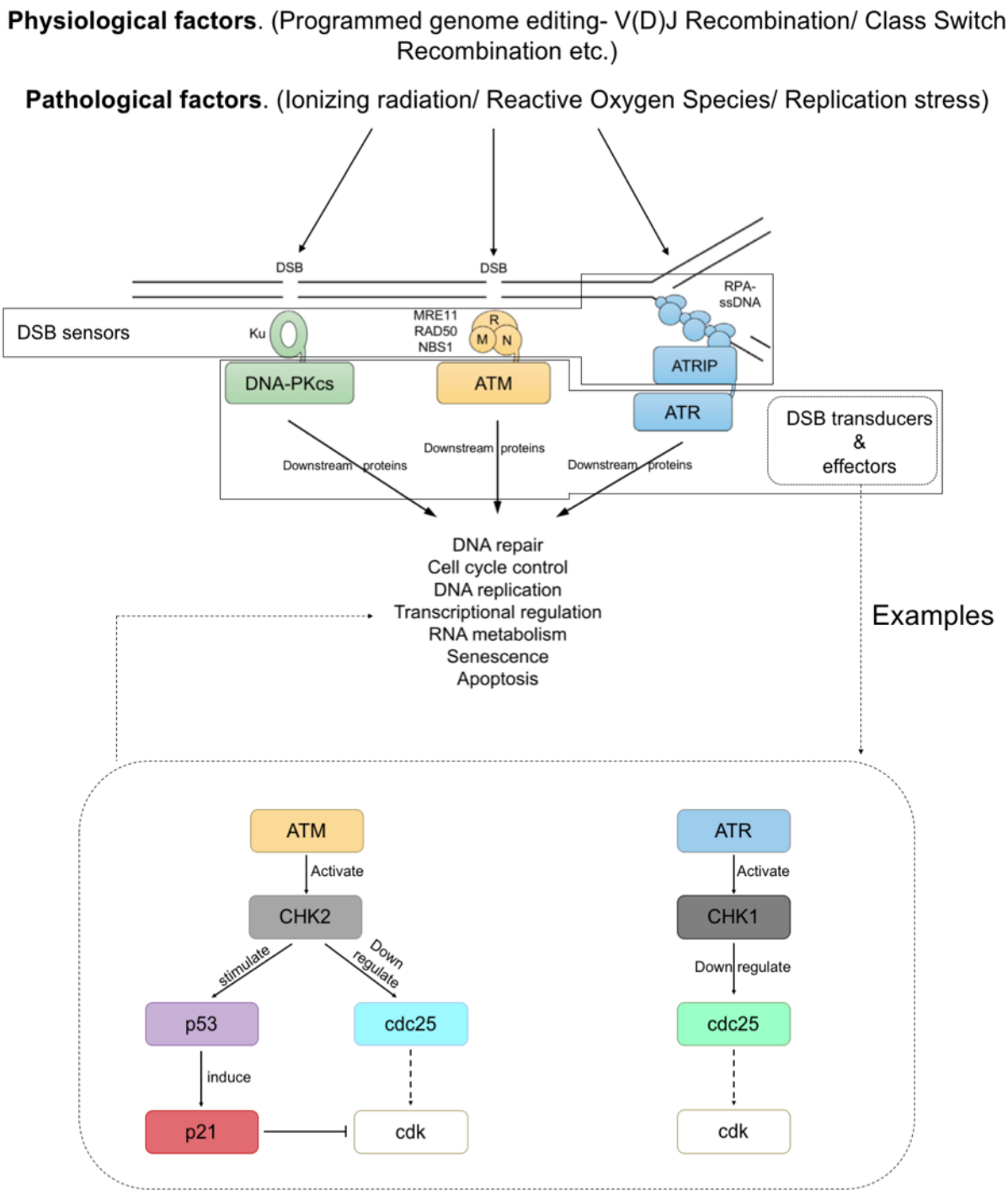


Figure 1. DNA double-stranded break damage response (Modified from Figure 2 of Blackford and Jackson 2017). Different factors can result in DNA DSBs and there are three main pathways that might respond including Ku/DNA-PKcs, MRN/ATM and RPA/ATRIP/ATR. Downstream effectors of the three pathways will lead to a series of responses including DNA repair, cell cycle control, transcriptional regulation, senescence and apoptosis.

DNA-PKcs, unlike ATM and ATR, is more directly involved in a DSB-repair pathway called non-homologous end joining (NHEJ). It is recruited to DSB sites via the interaction with Ku80 C-terminal domain (CTD) of Ku70/80 complex, which binds to double-strand DNA ends directly with a high binding affinity (Dvir *et al.*, 1992; Gottlieb and Jackson, 1993). Although the molecular mechanism of the activation of kinase activity needs further investigation, DNA-PKcs becomes activated when interacting with Ku70/80 and DNA, phosphorylating a series of proteins including many of the NHEJ components including itself. However, it remains unsure if the phosphorylation of those components except of itself is important for NHEJ.

Moreover, activated DNA-PKcs also phosphorylates a few of ATM/ATR substrates. For example, DNA-PKcs regulates the phosphorylation of H2AX and has an effect on cell-cycle progression (An *et al.*, 2010). However, the downstream DDR pathways of DNA-PKcs, except for DSB repair, are not as clear and extensively studied as those of ATM and ATR. It was recently shown that DNA-PKcs regulates the DNA damage response through inhibitory phosphorylation on ATM that impairs the signalling pathway of ATM upon DSB (Zhou *et al.*, 2017).

1.3 Human DSB Repair Pathways

To repair DSBs, there are two main mechanisms in our cells-- homologous recombination (HR) and non-homologous end joining (NHEJ) (Chang *et al.*, 2017). There are also two minor and intrinsically mutagenic pathways called alternative end joining (alt-ER) and single-strand annealing (SSA), which will not be discussed in detail in this thesis. As the name implies, a homologous template (sister chromatin) is needed for HR to make sure the DSB is repaired correctly using the other intact double-strand DNA. While in the case of NHEJ, the two ends of DSB are directly joined together without any template, it is important for the cell to fix the DSBs using the appropriate pathway and the decision of this process is not fully understood but there are several proposals as to how our cells make this decision.

At the cellular level, the pathway choice of DSB is highly connected to the cell cycle phase. NHEJ is functional throughout the whole cell cycle except for M phase, while HR is limited to the S/G2 phase when sister chromatin is available.

At the molecular level, there are two pathways— one by BRCA1/CtIP and the other by 53BP1 that promote HR (DNA end resection) and NHEJ (DNA end protection) respectively. When 53BP1 comes to the DSB site through the interaction with the modulated histone markers of the ubiquitylated K15 of H2A-type histones, it can be phosphorylated by kinases including ATM (Fradet-Turcotte *et al.*, 2013). RIF1 then binds to the phosphorylated 53BP1, suppressing the repositioning of 53BP1 and making the chromatin region too compact for other nucleases to come close (Chapman *et al.*, 2013). This limits the DNA end resection for HR and directs the pathways to NHEJ. BRCA1 can antagonize 53BP1 in different ways. For example, BRCA1 could promote a phosphatase called PP4C, which could further dephosphorylate 53BP1 and release the binding of RIF1 (Isono *et al.*, 2017). 53BP1 could then reposition on the chromatin, to allow nucleases including EXO1 to do the resection of DNA end for HR or other pathways that require DNA end resection including single-strand annealing (SSA) and alternative end joining (alt-EJ) (Ceccaldi, Rondinelli and D'Andrea, 2016). In addition to the marker of H2AX, recent research showed that the methylation of H4K20 also plays a role in DSB-pathway decision as H4K20me0 can be recognised by ARD of BARD1, which is in complex with BRCA1

(Nakamura *et al.*, 2019). However, 53BP1/RIF1 can bind to the site of H4K20me2, leading to the process of NHEJ (Botuyan *et al.*, 2006; Greeson *et al.*, 2008).

Also, the pathway choices can be connected with DNA-end types. When one-end DSBs are formed at the replication fork, HR is the main pathway to fix them. However, interestingly, in the G2 phase when two-end DSBs occur (i.e. caused by IR), NHEJ dominates the repair process, fixing around 70% of them, while HR takes charge of repairing the rest 30% despite the opinion that HR is more accurate and secure and NHEJ is more error prone and mutagenic (Shibata *et al.*, 2011). In fact, NHEJ is highly efficient and accurate under most of the cases (Bétermier, Bertrand and Lopez, 2014).

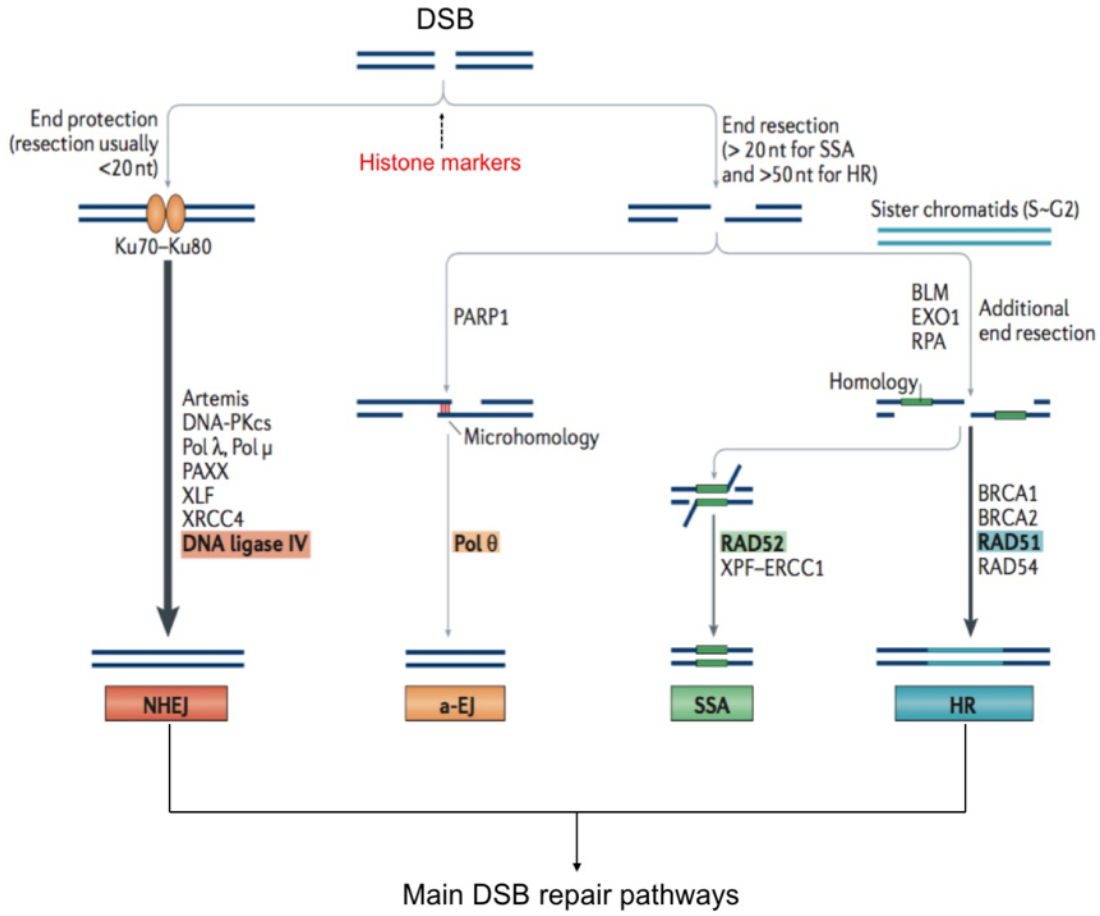


Figure 2. Human DNA DSB repair pathways (Modified from Figure 4 Chang et al., 2017). There are four DSB-repair pathways in total: non-homologous end joining (NHEJ), homologous recombination (HR), alternative end joining (a-EJ) and single-strand annealing (SSA). Among the four pathways, NHEJ and HR are the dominant pathways responsible for most DSB repair and will be further introduced in following chapters. SSA and a-EJ pathways are minor and mutagenic pathways that will not be discussed.

1.3.1 Homologous Recombination

Homologous recombination (HR) can be divided into three main steps- end resection, Rad 51 assembly and stand invasion and resolution of the intermediate. It starts with the DNA end resection to produce 3' single-strand DNA (ssDNA). This initial step involves many proteins, including MRN, CtIP, EXO1 and BLM (Sartori *et al.*, 2007; Nimonkar *et al.*, 2011). The trimmed 3' ssDNA can then become a good binding partner for Rad51, which is a DNA-dependent ATPase, to form nucleoprotein filament. However, the ssDNA is firstly bound by the replication protein A (RPA) (Zelensky, Kanaar and Wyman, 2014). To facilitate the formation of Rad51-DNA complex, many mediator proteins are required. Basically, BRCA1 recruits BRCA2 through the binding partner PALB2 (Xia *et al.*, 2006; Sy, Huen and Chen, 2009; Zhang *et al.*, 2009). BRCA2 then promotes the binding of Rad51 to ssDNA while removing RPA (Zelensky, Kanaar and Wyman, 2014). The defining step of HR is that 3′ Rad51-ssDNA complex will invade into a homologous duplex, forming the strand invasion intermediate (D loop) (Jasin and Rothstein, 2013). The invading 3′ ssDNA is used as a primer and extended by DNA polymerase to copy the missing information from the unbroken homologous duplex.

The D loop will later be resolved in several ways, which will not be introduced in detail here, leading to the final result of noncrossover (NCO) or crossover (CO) (Klein and Symington, 2012). Generally, the resolution of the intermediate, which may be regulated by the cell cycle (Matos *et al.*, 2011), is complicated, involving many proteins including BLM/TOP3α /RMI1 complex, MUS81/EME1 complex, GEN1 and SLX1/SLX4 (Wu and Hickson, 2003; Ho *et al.*, 2010; Wechsler, Newman and West, 2011; De Muyt *et al.*, 2012; Zakharyevich *et al.*, 2012)

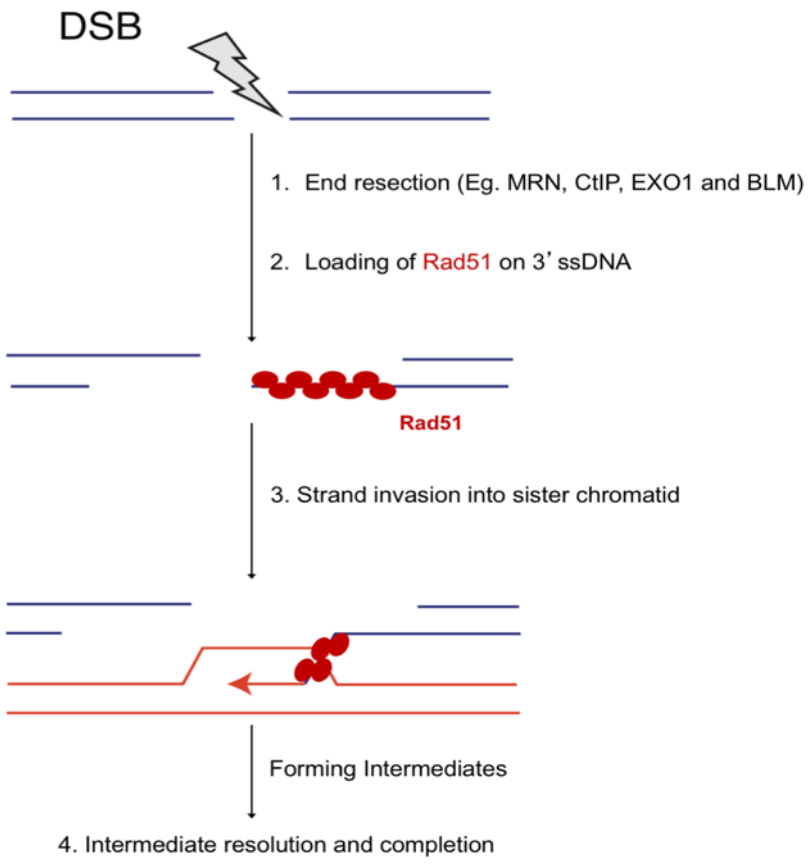


Figure 3. Schematic diagram of homologous recombination.

1.3.2 Non-Homologous End Joining

As mentioned, non-homologous end joining (NHEJ) is one of the main DSB repair pathways and the preferred repair pathway in humans and other higher-level organisms including vertebrates. It is highly dynamic and has many components used for different types of DSBs. However, no matter how complex the process of NHEJ may be, it can be summarised in terms of temporal organisation and categories of proteins involved.

The temporal organisation is usually separated into three mains steps: end recognition, end synapsis/bridging and processing, and end ligation (Liang *et al.*, 2017; Reid *et al.*, 2015; Ochi, Wu and Blundell, 2014).

The first step, end recognition, is the starting point where the DSB site is located. Ku70/80 first binds to the DSB ends and recruits DNA-PKcs via the C-terminal domain (CTD) of Ku80 to form the holoenzyme of DNA-PK (DNA, Ku70/80 and DNA-PKcs) (Lieber, 2010). DNA-PKcs then becomes activated and phosphorylates other downstream proteins and itself, also recruiting some other components including Artemis, which is one of the major nucleases involved in NHEJ (Chang and Lieber, 2016).

The second step, end synapsis/bridging and processing, is to make sure the DNA ends are spaciously closed together for further ligation when they are ligable. This is also the most dynamic and complex step. DNA-PK molecules interact with each other and can form synapsis (DeFazio *et al.*, 2002; Ma *et al.*, 2004; Spagnolo *et al.*, 2006). A recent single-molecule study has also demonstrated that PAXX can also be involved in the end synapsis together with DNA-PK (Wang *et al.*, 2018). Moreover, a crystallisation and EM study showed that XLF/XRCC4 can form elastic and flexible long filaments, which may be helpful for DNA bridging especially when the DSB ends are distant apart (Roy *et al.*, 2012a; Mahaney *et al.*, 2013). Sometimes, the ends are damaged or not ready for ligation. For example, there are usually many modifications including crosslinks or base modifications in the cases of DSBs caused by ionizing radiation. Another classical example is the V(D)J recombination process where the physiological DSBs are produced for the recombination of V/D/J segments (Roth, 2014). RAG1/2 complex comes to the flanking region of the targeted segments and cleaves the DNA

to create hairpin ends, which cannot be ligated and needs an endonuclease named Artemis to open the hairpin for ligation (Chang and Lieber, 2016).

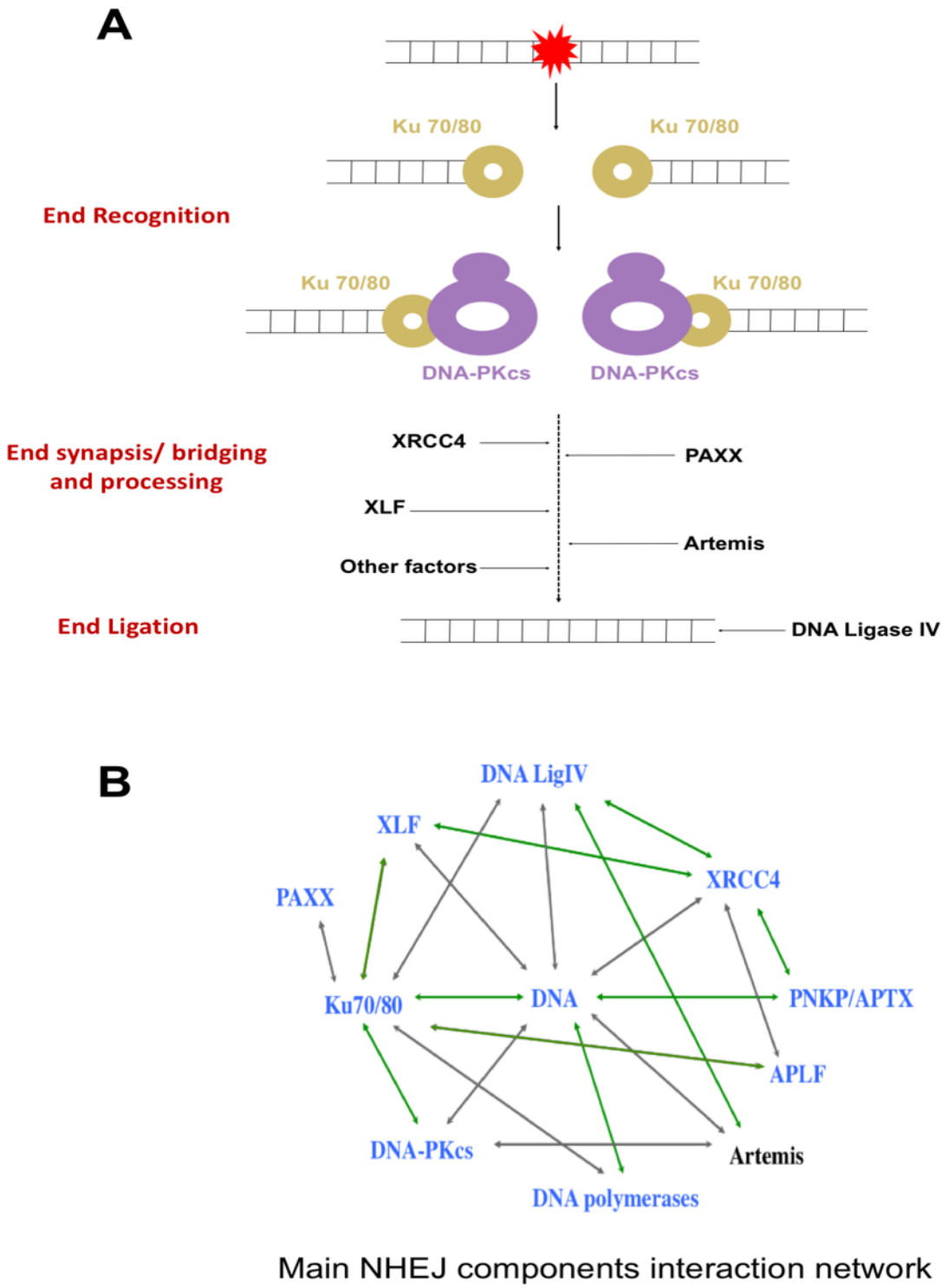


Figure 4. Non-homologous end joining (NHEJ) temporal and spatial organisation. (A) Temporal organisation of NHEJ: NHEJ can be divided into three steps—end recognition, end synapsis/bridging and processing, and end ligation; (B) Spatial organisation of NHEJ: Main components of NHEJ are labelled in the network. A blue label means that the structure of the protein or a part of the protein is known while a black label means unknown. An arrow indicates binary interaction. A green arrow means that the structure of the protein complex is known while grey arrow means that there is no structural information of the complex.

The third and final step of NHEJ, end ligation, involves joining the two broken ends, and mediated by the NHEJ ligase complex: DNA ligase IV, XRCC4, and XLF. XLF and XRCC4 can affect the activity and stability and more details will be introduced later in the thesis. It is also possible that the newly discovered protein PAXX is playing a role in the NHEJ, as it is structurally homologous to XRCC4 and XLF (Ochi *et al.*, 2015). It is kown that PAXX functions with XLF and XRCC4 to regulate the progress of NHEJ. Moreover, PAXX can increase the synapsis duration by 7 times when added to the Ku/DNA-PKcs/XRCC4/XLF/Ligase IV reaction system (Wang *et al*., 2018).

From the perspective of the categories of proteins involved in human NHEJ, there are five classes of proteins: the DNA-PK complex, the nucleases, the polymerases, the ligase complex and others. The DNA-PK complex and ligase complex were mentioned previously in the NHEJ temporal organisation and as they are indispensable for the process while the other kinds of protein may be involved in NHEJ under different conditions.

The nucleases involved in NHEJ mostly are helping to resection the 5’ or 3’ overhang by the exonuclease or endonuclease when the ends are incompatible. This will expose or generate small regions of nucleotides (usually 2-4) with microhomology between two DNA strands to facilitate the end ligation. As mentioned previously, Artemis is the major nuclease involved in NHEJ. Around 20%- 50% of the DSBs caused by ionizing radiation need Artemis for repair (Chang and Lieber, 2016). Other nucleases that may contribute in this process include but not be limited to the aprataxin and PNKP-like factor (APLF), ERCC1, ERCC4 (XPF), and exonuclease 1 (EXO1).

The polymerases participating in NHEJ have 4 members: Pol λ, Pol μ, TdT, Pol θ. They all belong to a subfamily of DNA polymerases: Pol X family polymerases. Although it is now recognised that NHEJ is actually efficient and generally accurate, the polymerases can introduce extensive mutations.

Other NHEJ components are required in NHEJ mainly due to the need to modify the DNA ends. For example, polynucleotide kinase (PNKP) is needed when the 5’ end is missing the phosphate group and requires phosphorylation. In human, PNKP is also a phosphatase. It

could also remove phosphate on the 3' ends that are caused by some oxidative damages (Bernstein *et al.*, 2005). Sometimes, DNA ligase IV does not complete the ligation and forms an intermediate product where an AMP group is still covalently bound to the 5'-end of one strand of the DSBs. Under this circumstance, an enzyme called aprataxin is needed to cleave the AMP group (Ahel *et al.*, 2006). Moreover, Tyrosyl DNA phosphodiesterase 1 (TDP1) is the only enzyme that is currently known to process the process 3'-phosphoglycolates (3'-PGs) (Kawale and Povirk, 2018). 3'-PGs is a common damage of DSB caused by ionizing radiation, accounting for 10% of all the IR-induced DSB (Zhou *et al.*, 2005).

Although NHEJ is complicated with numerous components involved in different interacting/processing steps, certain NHEJ components are central to the process and relevant to the background of this PhD thesis. I will therefore give an introduction to these components in detail.

1.3.2.1 Ku70/Ku80

Ku70/80 is a protein complex composed of two proteins, Ku70 (69kDa) and Ku80 (83kDa), which have similar topology and form together a ring structure that binds DNA double helices at double-strand breaks (Figure 5).

Each protomer has a N-terminal α/β domain, belonging to the von Willebrand (vWA) family, which is made up of a six-stranded β sheet in a Rossman fold (residues 35-251 in Ku70; residues 8-165 in Ku80). This is followed by a ß-barrel central domain (residues 257-436 in Ku70; residues 199-423 in Ku80) and an ARM domain (residues 440-528 in Ku70; residues 427-542 in Ku80) (Walker, Corpina and Goldberg, 2001). In the case of Ku70, the globular SAP (SAF-A/B, Acinus and PIAS) domain (559-609) forms a three-helical bundle at the C terminus (Zhang *et al.*, 2001). The C-terminal region of Ku80 has a linker between the N-terminal globular domain/ core domain and a C-terminal 6-helical bundle (595-704), which is followed by a conserved region at the C terminus (721-732) that interacts as a helix with DNA-PKcs recruiting it to the DNA damage site (Zhang *et al.*, 2004).

Ku70/80 is one of the most important components in NHEJ and has homologues in yeast and some bacteria (Matthews and Simmons, 2014; Emerson and Bertuch, 2016). The homologues in bacteria are much smaller at 30-40 kDa than Ku70/80. In addition, unlike Ku70/80, the bacterial homologues have only the central domain, which is mainly the heterodimerization region, without the vWA domain and C-terminal domain (Pitcher, Brissett and Doherty, 2007).

In the Ku70/80 complex, the vWA, the ß-barrel central and the ARM domains together form the globular domains of the complex. The heterodimerisation of the complex is mediated by the central domain, of which the extended loops and α helices form the interconnected ß-strands (ARM). The ring structure has a positively-charged inner surface that binds DNA of minimum length 14bp (Walker, Corpina and Goldberg, 2001). Ku70/80 binds to DNA double-strand break ends strongly with affinity in the nM range. Although it was proposed that the SAP domain may play a role in the interaction with DNA, the C-terminal domain does not contribute to the binding affinity to DNA (Frit *et al.*, 2019). Furthermore, there is no sequence specificity of Ku-DNA binding. Although it does not bind to some specific DNAs including circular DNA, linear single-strand DNA and supercoil DNA, Ku70/80 can bind to various DNA

double-strand ends including blunt ends, DNA ends with short 5′ and 3′ overhang, DNA hairpin, cisplatin and damaged DNA ends caused by ionising radiation (Ono *et al.*, 1994; Turchi and Henkels, 1996; Pang *et al.*, 1997; Walker, Corpina and Goldberg, 2001; Arosio *et al.*, 2002).

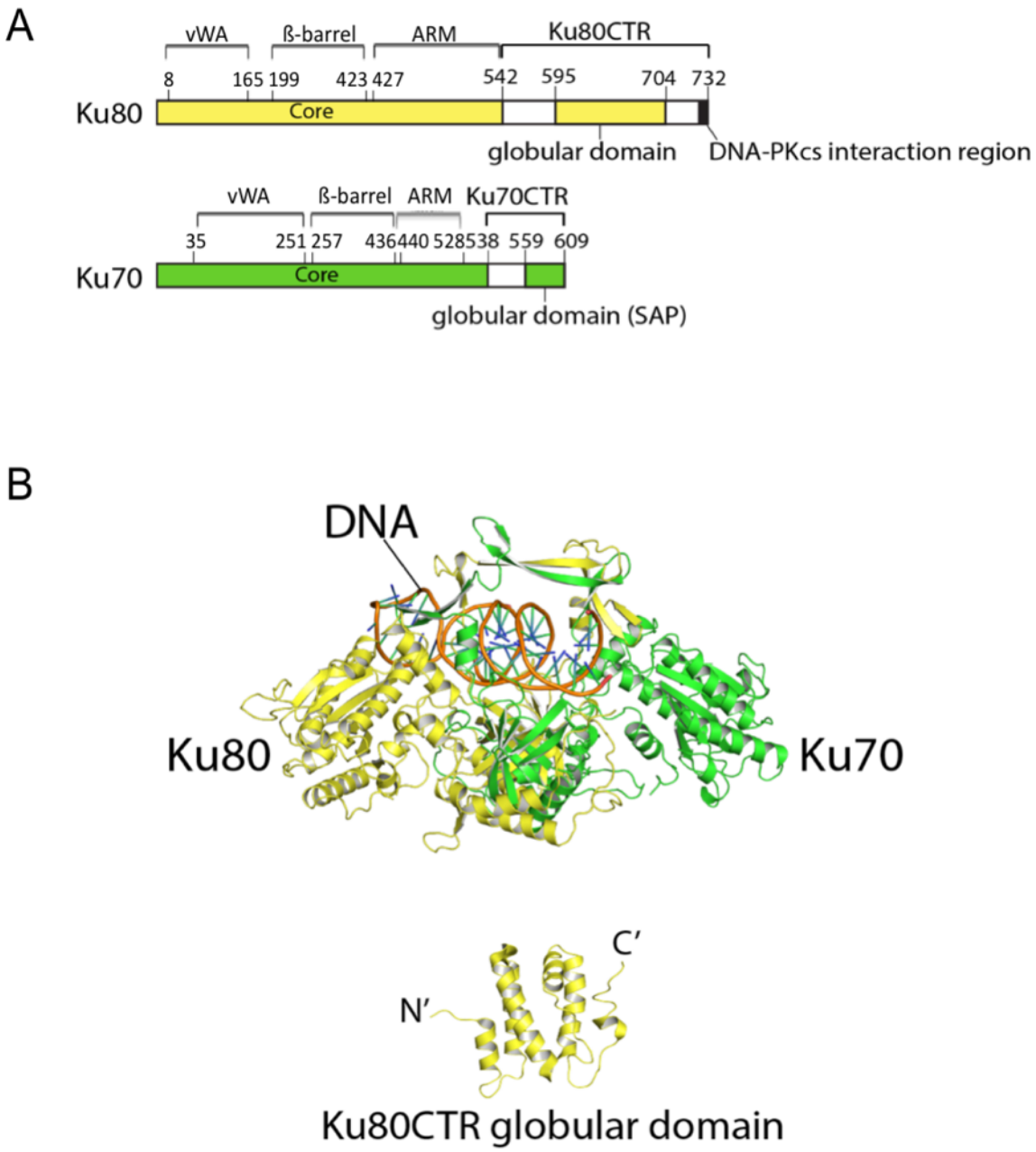


Figure 5. Structure of Ku70/80 complex (Modified from Figure 1 of Wu et al., 2019). [A] Schematic diagram of Ku70 and Ku80- Both Ku70 and Ku80 have a similar structural organisation of a core region (including vWA domain, ß-barrel domain and the ARM domain) and the C-terminal region (CTR); [B] Crystal structure of Ku70/80 core region in complex with DNA (Walker et al., 2001) and the NMR structure of the Ku80 CTR globular domain (Zhang et al., 2004).

The roles of the two vWA domains are not fully understood. The Ku80 vWA domain plays a role in the binding of many other Ku-binding proteins, containing the Ku-binding motifs (KBM) including APLF, XLF, WRN and CYREN (Li and Comai, 2000, 2001; Grundy *et al.*, 2012, 2016; Rulten and Grundy, 2017; Nemoz *et al.*, 2018). These protein-protein interactions play important physiological role. The binding of Ku with APLF and XLF can promote the accuracy and efficiency of NHEJ (Nemoz *et al.*, 2018). The interaction between Ku80 and WRN regulate the pathway choice between NHEJ and alt-EJ (Shamanna *et al.*, 2016). CYREN (cell-cycle regulator of NHEJ) is an NHEJ inhibitor at the deprotected telomeres to prevent chromosome joining in the S and G2 phases (Arnoult *et al.*, 2017). Although no structure is available, it was shown that the N-terminal KBM of CYREN binds to Ku and can be displaced from Ku by an excess of APLF A-KBM, indicating that CYREN is binding to the same site of Ku80 vWA (Grundy *et al.*, 2016). The interaction between CYREN N-terminal KBM and Ku80 inhibits NHEJ. The function of Ku70 vWA domain, on the other hand, is not as well defined as that of Ku80. Mutated Ku70 S155/D156 helps to increase the survival rate following the IR treatment; in fact the S155A mutation on its own is sufficient to increase the survival rate while the phosphomimetic mutation (S155D) was able to reverse the effect via controlling the activation of transcriptional factor 2 (ATF2) and the downstream apoptosis pathway (Fell and Schild-Poulter, 2012). This indicates that the Ku70 vWA domain may play a regulatory role in signalling of the DDB but the mechanism remains unclear.

The role of the C-terminal domains (CTD) also remains unclear. Ku70 CTD has been proposed to be involved in DNA binding both through the flexible linker region (536–560) and the last 10 amino acid residues (600-609) (Chou *et al.*, 1992; Wang, Dong and Reeves, 1998). In the case of Ku80, the NMR structure of the six-helical bundle of C-terminal domain showed that the conserved residues clustered on two major surface regions, indicating that this domain is involved in protein-protein interaction (Zhang *et al.*, 2004). However, 15 years after the structure was solved, the interaction partner still remains unclear as is the function of the six-helical bundle. The last helix of 12 amino-acid residues at the C terminus of Ku80 interacts specifically with DNA-PKcs and plays a role in activating the DNA-PKcs kinase activity, which will be further described in the following section on DNA-PKcs (Singleton *et al.*, 1997; Gell and Jackson, 1999).

An important question about the Ku70/80 complex is how and when it binds to and moves away from DNA ends. Ku is known to bind strongly to the DNA ends on its own, providing the first step in NHEJ. PARP1 may be a regulator of Ku binding to DNA as Ku70/80 dominates the DSB ends in the G1 phase, while PARP1 later competes with Ku70/80 in the S and G2 phases (Yang *et al.*, 2018). Moreover, through its enzymatic activity, PARP1 removes Ku70/80 in S/G2 phase (Yang *et al.,* 2018). Ku is recruited to the DSB sites rapidly but only stays there for a short period of time as the Ku signal decreases steadily a few hours after the initial damage in laser microirradiation experiments (Kim *et al.*, 2005; Mari *et al.*, 2006). It is unlikely that the Ku70 and Ku80 dissociate from DNA due to the high affinity of binding and the structure of Ku70/80 in complex with DNA does not reveal any possible releasing conformational change (Walker, Corpina and Goldberg, 2001; Rivera-Calzada *et al.*, 2007). Study on *Xenopus laevis* showed that the polyubiquitination of Lys48 of Ku80 will lead to the degradation by SCF (Skp1-Cul1-Fbxl12) E3 ubiquitin ligase complex (Postow *et al.*, 2008). It was also found in human that E3 ubiquitin ligase RNF8 (RING finger protein 8) would result in the polyubiquitination of Ku80 (Feng and Chen, 2012). The depletion of RNF8 will decrease the efficiency of NHEJ and extend the Ku retention on DNA (Feng and Chen, 2012). There is also another hypothesis that DNA direct nicking allows the escape of Ku70/80. In the yeast system, MRX (Mre11- Rad50-Xrs2) can perform an endonucleolytic incision next to the DNA end followed by DNA digestion to allow Ku70/80 releasing (Neale, Pan and Keeney, 2005; Wu, Topper and Wilson, 2008; Langerak *et al.*, 2011). These proposed mechanisms are quite different and more investigation is needed. To point out, the Ku70/80 removal studies were carried out in different organisms and those different mechanisms could be due to the divergence between yeast and higher eukaryotes just like many components and interaction of NHEJ.

1.3.2.2 DNA-PKcs/ DNA-PK

DNA-PKcs (DNA-dependent protein kinase catalytic subunit), as mentioned previously, is the core component of DDR and NHEJ. It is the largest member of the family of PI3-kinases with 4128 amino-acid residues, mainly composed of four parts: HEAT repeats (Huntingtin, Elongation Factor 3, PP2 A, and TOR1), FAT region [FRAP (FKBP12-rapamycin– associated protein), ATM (ataxia-telangiectasia mu- tated), TRRAP (transformation/transcription domain associated protein)], the kinase region and the FATC (FAT C-terminal) region.

Although the DDR pathways of DNA-PKcs are not fully understood, *in vitro* studies have shown that DNA-PKcs can phosphorylate a series of NHEJ components including Ku70/80 (Chan *et al.*, 1999; Douglas *et al.*, 2005), Artemis (Goodarzi *et al.*, 2006), XRCC4 (Yu *et al.*, 2003), XLF (YU *et al.*, 2008), PNKP (Zolner *et al.*, 2011), and DNA-PKcs itself (Douglas *et al.*, 2002, 2007; Meek *et al.*, 2007). Interestingly, although DNA-PKcs belongs to the family of PI3-kinases, these *in vitro* experiments show that the phosphorylation pattern of DNA-PKcs does not have the expected SQ/TQ consensus. The interaction of DNA-PKcs and Artemis activates the endonuclease activity of Artemis, which is indispensable for V(D)J recombination. However, whether this is due to phosphorylation of Artemis by DNA-PKcs remains unclear. Actually, in the aforementioned cases (Artemis, PNKP, XRCC4 and XLF), the phosphorylation after DDR is conducted more by ATM rather than or in addition to DNA-PKcs (Jette and Lees-Miller, 2015). The crosstalk between ATM and DNA-PKcs on many NHEJ proteins is complicated. Nevertheless, there is one well-understood target of DNA-PKcs phosphorylation and it is DNA-PKcs itself.

The autophosphorylation of DNA-PKcs mainly sits at two clusters-- ABCDE cluster (2609- 2647; T2609, S2612, T2620, S2624, T2638, T2647) and PQR cluster (2023- 2056) (S2023, S2029, S2041, S2053, S2056). *In vitro* experiments showed that DNA-PKcs can have autophosphorylation at many SQ/TQ sites (T2609, S2612, T2638 and T2647) and non SQ/TQ sites (S2624, S3205) (Douglas *et al.*, 2002). *In vivo* experiments showed that autophosphorylation occurs on even more residues (S2056, T2609, S2612, T2620, S2624, T2638, T2647, T3950) and the phosphorylation on S2056 after DSB is widely recognised as a reliable indicator of DNA-PK activation (Cui *et al.*, 2005; Douglas *et al.*, 2007; Meek *et al.*, 2007). ATM/ATR also phosphorylate on the ABCDE cluster. ABCDE and PQR clusters

reciprocally mediate end processing of NHEJ (Cui *et al.*, 2005). Phosphorylation on the ABCDE cluster promotes end processing while phosphorylation on the PQR cluster inhibits it. Also, Phosphorylation on the ABCDE cluster helps but is not sufficient or necessary for the kinase dissociation (Douglas *et al.*, 2007; Uematsu *et al.*, 2007; Dobbs, Tainer and Lees-Miller, 2010). Without proper structural information of DNA-PKcs and the structure of DNA-PKcs in complex with binding partners including Ku70/80 and Artemis at the atomic level, it is difficult to understand how the mechanisms of the autophosphorylation affect the complicated physiological roles DNA-PKcs plays in NHEJ.

Structural studies on DNA-PKcs have been challenging due to the size of this protein but were first carried out over 20 years ago using electron microscopy (EM) and atomic force microscopy (AFM) (Cary *et al.*, 1997; Yaneva, Kowalewski and Lieber, 1997). The structure of DNA-PKcs, the catalytic subunit of DNA-PK was reported at 22Å resolution using the method of cryo-EM imaging and electron crystallography (Chiu *et al.*, 1998; Leuther *et al.*, 1999). Later, EM studies showed that DNA-PKcs is a circular structure with a crown/head domain with the highest resolution at 7Å (Boskovic *et al.*, 2003; Rivera-Calzada *et al.*, 2005, 2007; Williams *et al.*, 2008). cryo-EM was also used to study the holoenzyme of DNA-PK (Spagnolo *et al.*, 2006).

Our group published the first crystal structure of DNA-PKcs in complex with Ku80CTD at 6.6 Å resolution but the N-terminal region was poorly defined and not modelled (Sibanda, Chirgadze and Blundell, 2010). It has been difficult to model the DNA-PKcs not only because of the low resolution and flexibility but also because it is difficult to obtain the correct register of the sequence due to the repetitive nature of the structures of the HEAT repeats.

The breakthrough in the structure determination of DNA-PKcs came in 2017 when a subatomic resolution was reported at 4.3 Å of the DNA-PKcs in complex with Ku80 CTD (Ku80 539-732), in which all the secondary elements and subdomains of the four main regions could be visualised (Chirgadze *et al.*, 2017; Sibanda *et al.*, 2017). To solve the sequence registration problem caused by the HEAT repeats, the method of selenomethionine (Se-Met) labelling was used. There were 228 Se-Mets evenly distributed in the two molecules of the asymmetric unit to help identify the sequence. However, only 90% of all 4128 residues were modelled into the electron density of the crystal structure.

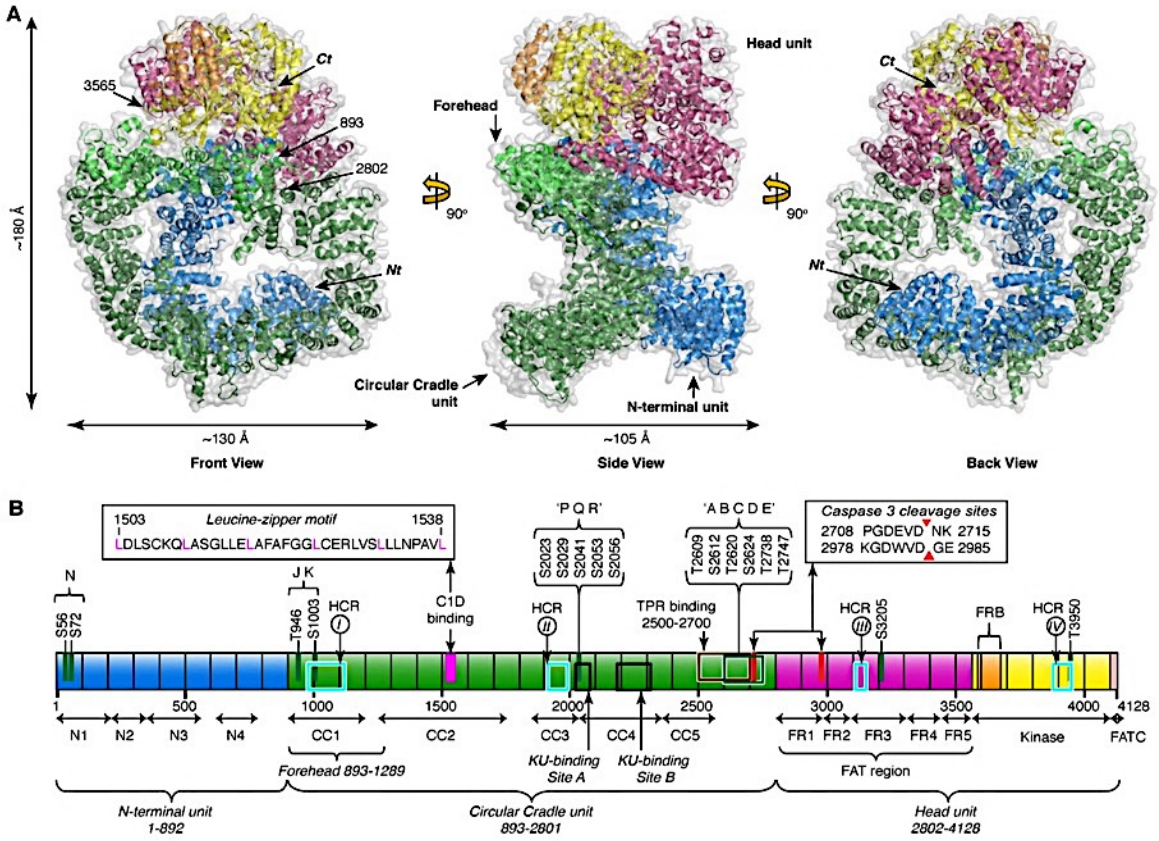


Figure 6. **Structural information of apo DNA-PKcs** (Adapted from Figure 1 of Sibanda et al., 2017). [A] DNA-PKcs molecule from the crystal structure of DNA-PKcs in complex with Ku80CTD with different structural units labelled with different colours (PDB code: 5LUQ). The N-terminal region is coloured blue; Circular Cradle is coloured green; FAT region is coloured pink; Kinase region is coloured yellow and FATC region is coloured light pink. [B] Schematic diagram of DNA-PKcs structural composition of the three big units (N-terminal, Circular Cradle and Head) and the supersecondary structural units.

Structurally, DNA-PKcs could be separated into three parts: the N-terminal region (1-892), the Circular Cradle (893-2801) and the C-terminal Head (2802-4128). The N-terminal region has 38 α-helices forming 4 supersecondary structures with continuous hydrophobic core (N1-N4). The Circular Cradle is the largest structural domain with 85 α-helices forming 5 supersecondary structures (CC1- CC5). The C-terminal Head unit, containing the FAT domain, kinase domain and FATC domain, has 64 α-helices. However, the region 2576-2744 was difficult to model with four-α-helices (probably representing 2602-2665) hanging down in the centre of the empty cavity of DNA-PKcs. This is due to missing or weak densities of this region making it difficult to model and define sequence registration. This implies that the region is flexible and not rigidly structured. Interestingly, this missing region covers the ABCDE cluster, which is important in the autophosphorylation and DNA-PKcs regulation. It is surprising that

this region is not structurally stable considering its physiological importance. This region could possibly be interacting with other binding partners and more detailed structural information, possibly at higher resolution, will be helpful to explain its role in the regulation of DNA-PKcs. In addition to the missing density of DNA-PKcs, the density of Ku80 CTD (Ku80 539-732) was largely missing. Only three α-helices were modelled in the structure. The helix of the C-terminus of Ku80 is confirmed to be located at the tip of the circular cradle sitting close to the PQR cluster while the other helix-loop-helix remains unknown. The globular domain (Ku80 595-704) that had been previously studied using NMR was not detected and not built in the model.

Soon after the atomic model of DNA-PKcs of our crystal structure was published, cryo-EM structures of DNA-PKcs and DNA-PK from two independent groups were released at low resolution with chain tracing based on our crystal structure model (Sharif *et al.*, 2017; Yin *et al.*, 2017a). Although both of the cryo-EM studies shared the problem of preferred orientation of the sample, which leads to an anisotropic map and requires further optimisation, it showed how fast cryo-EM can build on and corroborate X-ray studies so improving our knowledge of the structures of large proteins and their complexes.

One group published the cryo-EM structure of apo DNA-PKcs at 4.4 Å resolution and DNA-PKcs with Ku70/80 at a resolution of 5.8 Å (Sharif *et al.*, 2017). In the apo DNA-PKcs structure, most of the structure is consistent with the crystal structure but the N-terminal region is lifted relatively in the cryo-EM structure. This may be due to the effect of crystal packing where the movement of different domains is more restricted in a crystal but freed in the solution environment of cryo-EM. The region (2575- 2775) containing the ABCDE cluster is not detected in cryo-EM either and there was even no EM density for the four-α-helices region (Ku80 2602-2665) previously modelled. This further indicates that the region may be rather flexible and needs more structural information. In the 5.8 Å resolution structure of DNA-PK cryo-EM structure, although the bands of DNA-PKcs, Ku70 and Ku80 were shown on SDS-PAGE gel, neither Ku70/80 nor DNA were detected in the final map. Instead, there is extra density located between the N-terminal region and the circular cradle HEAT repeats and it is proposed to be the globular region (Ku80 595-704) in the middle of Ku80 CTD. This can also explain why

the core domain of Ku70/80 is missing as there is a long flexible linking region (542-595) between the globular region and the core domain.

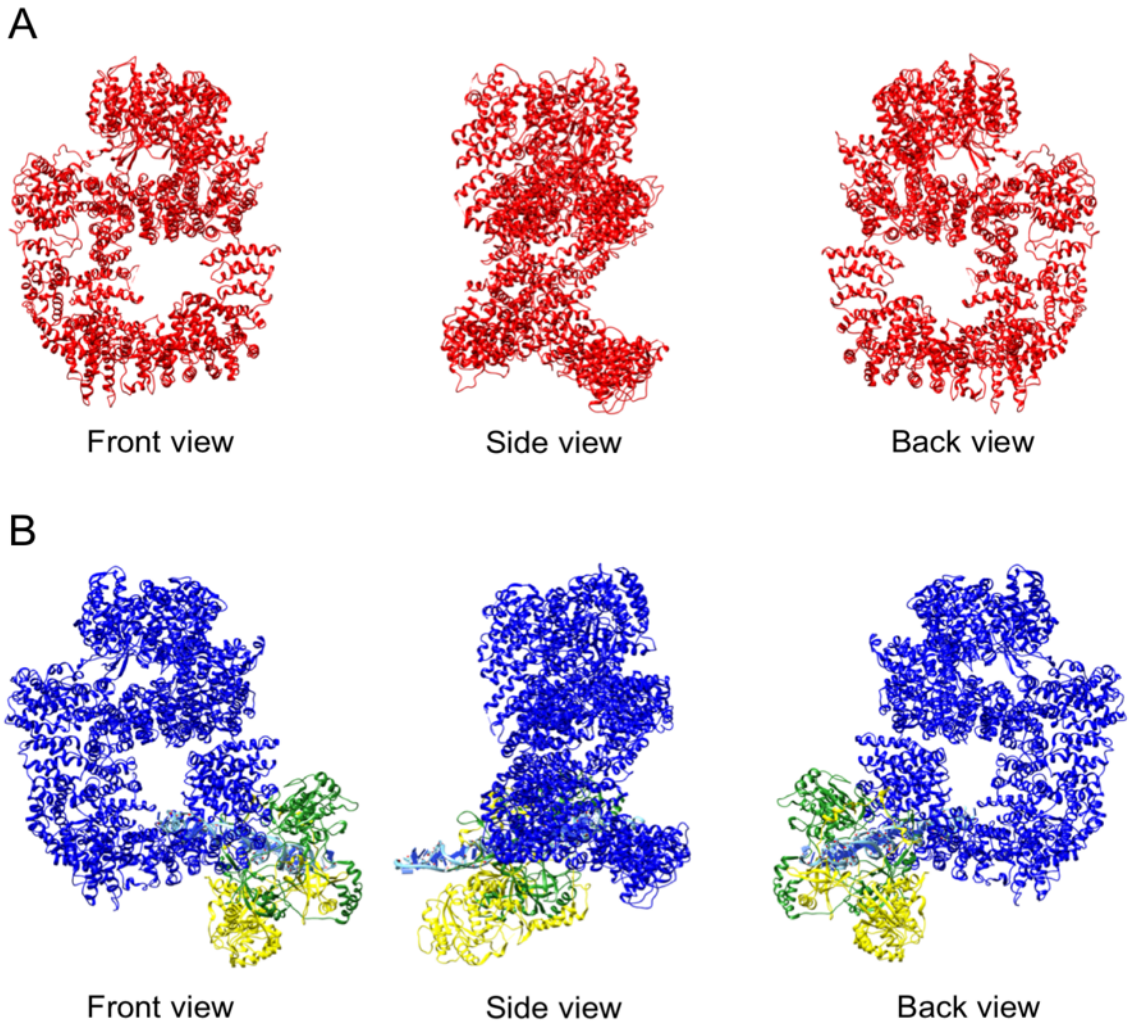


Figure 7. **Cryo-EM models of DNA-PKcs and DNA-PK**. [A] Cryo-EM model of DNA-PKcs in three different views including front, side and back view (PDB code: 5W1R). DNA-PKcs is labelled red [B] Cryo-EM model of DNA-PK in three different views including front, side and back view (PDB code: 5Y3R). DNA-PKcs is labelled blue; Ku70 is labelled green; Ku80 is labelled yellow and the dsDNA is labelled cyan.

The other group published the cryo-EM structure of DNA-PK at 6.6 Å resolution (Yin *et al.*, 2017a). Although at lower resolution, the extra density of DNA and Ku70/80 core domain are all present and the map shows the assembly of the holoenzyme—DNA goes through the ring of Ku70/80 and the N-terminal region of DNA-PKcs while Ku70/70 also sits next to the N-terminal region and circular cradle. The N-terminal region has a significant uplift compared to the apo form of DNA-PKcs. There is an alanine helix on the circular cradle of DNA-PKcs and it is proposed to be the helix from Ku80 CTD while the exact region on Ku remains unknown

due to the low resolution. In addition, the globular region of Ku80 CTD is missing in the holoenzyme structure unlike the previous cryo-EM map of DNA-PK. Last but not least, the ABCDE cluster included region (DNA-PKcs 2577-2773) is still missing.

Together, the three structures can give us an idea about how Ku interacts with DNA-PKcs (Sharif *et al.*, 2017; Sibanda *et al.*, 2017; Yin *et al.*, 2017). It is clear that the interaction between Ku and DNA-PKcs is not a simple Ku80 C-terminus/DNA-PKcs single-site interaction and a model of Ku-DNA-PKcs multistep interactions can be deduced from these structures. To start with, Ku interacts with DNA and recruits DNA-PKcs at a distance through the C-terminal 12 amino acids. Once recruited by the Ku80 C terminus, the globular region of Ku80 CTD will further interact with DNA-PKcs bewteen its N-terminal region and HEAT repeat, meanwhile releasing the C terminus at the PQR site, to bring DNA-PKcs closer to the Ku-DNA hub. This interaction may not be specific and strong but it moves DNA-PKcs very close to the DNA end. Finally, the DNA end comes in to replace the globular region and the core domain of Ku70/80 docks at the site between the N-terminal region and circular cradle. In the mean time, another Ku 80CTD region comes in and interacts with the circular cradle to prevent the DNA end from sliding.

The molecular mechanism for the activation of the kinase moiety of DNA-PKcs remains unclear with two main hypotheses about the mechanism. One hypothesis emphasises the interaction between K881 and E3933, which are close to the highly conserved region I (HCRI) (DNA-PKcs 979- 1121). This interaction is important for the activation as it opens up the active site for substrate processing. The other hypothesis is that, due to the significant uplift, the N-terminal region of DNA-PKcs interacts with the FAT domain, resulting in the opening up of the catalytic site. However, the current resolution of the structures may need further improvement to allow qualitative comparison of the conformational changes. Also, structures of intermediate state of catalysis (e.g. with an inhibitor) may be helpful for further understanding the kinase activation mechanism of DNA-PKcs.

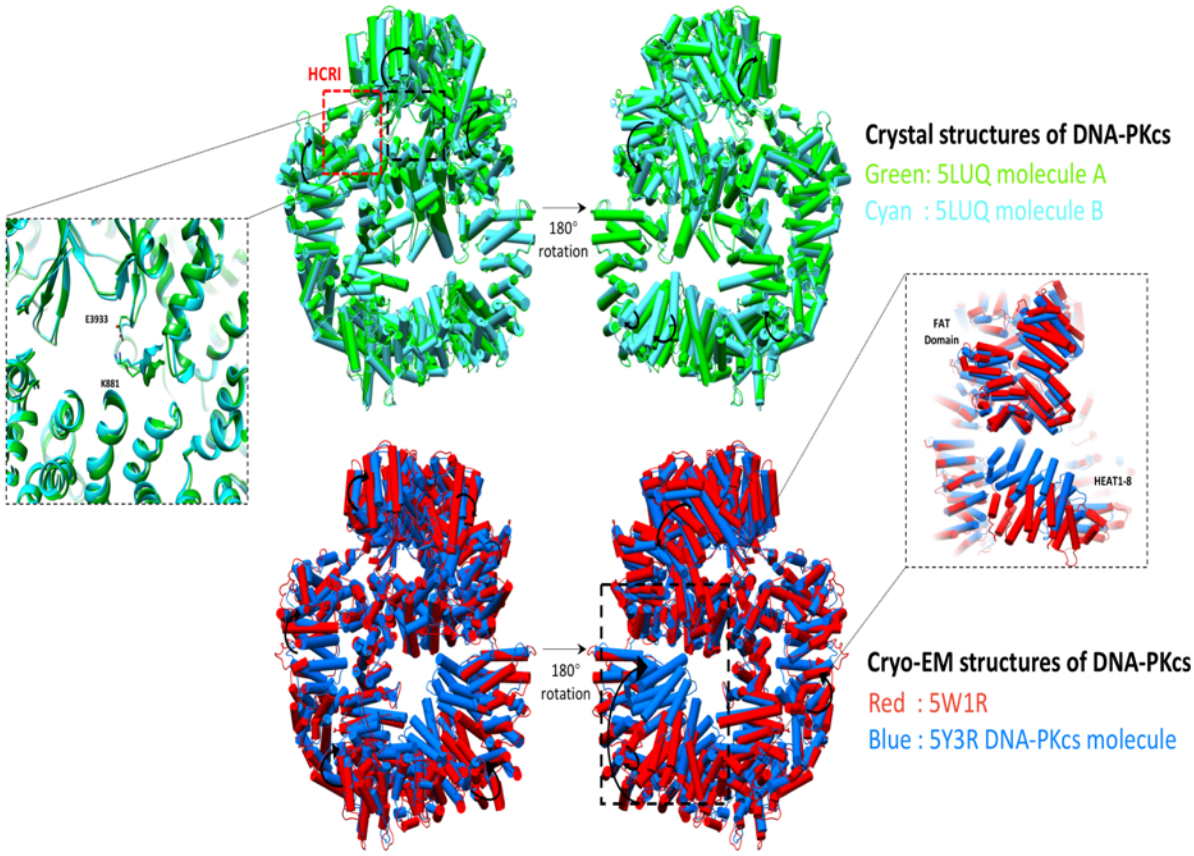


Figure 8. Allosteric activation of DNA-PKcs kinase activity (Adapted from Wu et al., 2019). Four DNA-PKcs models are used. In the top panel, two DNA-PKcs molecules (green and cyan) from crystal structure of DNA-PKcs in complex with Ku80CTD (PDB code: 5LUQ) are superposed. There is a small conformational change (Sibanda et al., 2017). The Highly Conserved Region I (HCRI), including the interaction between E3933 and K881, is amplified in the window on the left. In the bottom panel, the DNA-PKcs model from the cryo-EM structure of the apo DNA-PKcs (PDB code: 5W1R) (Sharif et al., 2017) is coloured red, superposed on the DNA-PKcs model (coloured blue) from the cryo-EM structure of DNAPK holoenzyme (PDB code: 5Y3R) (Yin *et al.*, 2017). The uplift of the N-terminal unit leads to interaction between the FAT domain and HEAT 1-8. This is amplified in the right window. To clarify, the Ku C-terminal helices are omitted from (Sibanda *et al.*, 2017) and DNA and Ku molecules are omitted from (Yin *et al.*, 2017) structures. The arrows indicate the conformational changes between the compared models.

1.3.2.3 XRCC4 Superfamily- XRCC4/XLF/PAXX

The XRCC4 superfamily members, XRCC4, XLF and PAXX, are paralogues with divergently evolved structures. Although there has been no enzymatic function reported in this superfamily, all three members participate in NHEJ and play important roles, some of which may be redundant. The first discovered member, after which the superfamily is named, is XRCC4. It is an indispensable core member of NHEJ (Giaccia *et al.*, 1990; Li *et al.*, 1995; Andres *et al.*, 2012). Mutations in XRCC4 can result in embryonic lethality in mice and immunodeficiency and developmental inhibitions in human (Andres *et al.*, 2012). Also, some mutations in XRCC4 are found to cause increased risk of cancer (Shao *et al.*, 2013). XLF (XRCC4 like factor) is the second member to be discovered (Buck *et al.*, 2006). Mutations in XLF can lead to radiosensitivity, immunodeficiency, neurodevelopmental disorder (microcephaly) and lymphopenia (Buck *et al.*, 2006). Shown to interacts with DNA Ligase IV complex, XLF promotes NHEJ (Ahnesorg, Smith and Jackson, 2006). PAXX is the last discovered member of the family and has been shown to promote NHEJ via interaction with key NHEJ components including Ku70/80 (Ochi *et al.*, 2015; Xing *et al.*, 2015; Craxton *et al.*, 2018).

XRCC4 is a 38kDa protein with 334 amino acids, which can be divided into 3 domains: the head domain (1-118), the helical tail domain (119-213) and the C-terminal domain (214-334) (Li *et al.*, 1995). It is a homodimeric protein which can also form tetramers *in vivo* (Mizuta *et al.*, 1997; Junop *et al.*, 2000; Modesti *et al.*, 2003;). The homodimer interacts with DNA Ligase IV to form the tight DNA Ligase IV complex (Sibanda *et al.*, 2001). So far there is no enzymatic function of XRCC4 reported.

The structural information for XRCC4 is available for the sequence before the C-terminal domain. The head domain has a ß-sandwich globular structure with a helix -turn-helix (HTH) motif (Junop *et al.*, 2000; Sibanda *et al.*, 2001). This head domain is important for the stabilisation of XRCC4 and for the interaction with XLF, which will be introduced later (Mizuta *et al.*, 1997; Grawunder *et al.*, 1998; Modesti, Hesse and Gellert, 1999). The head domain is then followed by a long helical tail that forms the supersecondary structure of coiled coil in the homodimer through residues 119-155 (Sibanda *et al.*, 2001).

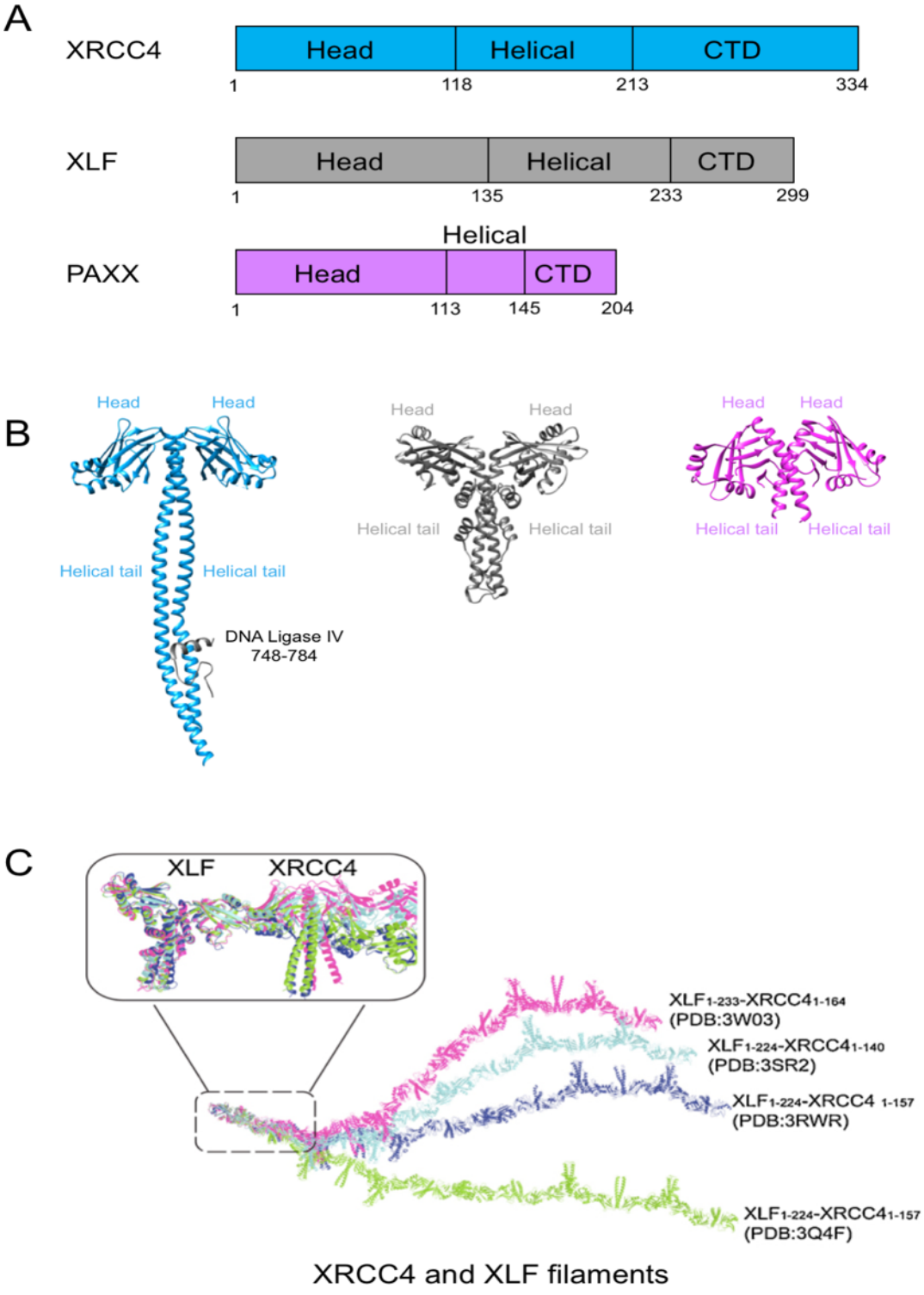


Figure 9. **XRCC4 superfamily: XRCC4, XLF & PAXX**. [A] Schematic diagram of XRCC4 (blue), XLF (grey) and PAXX (purple) showing similar structural composition of Head domain, Helical tail and C-terminal domain (CTD); [B] Structures of XRCC4 superfamily: XRCC4 (residues 1-213; blue) in complex with DNA Ligase IV peptide (residues 748-784; Dark grey); XLF (residues 1-233; grey); and PAXX (residues 1-204; purple). To point out, there is no structural information of CTD of any XRCC4 superfamily member; [C] Structures of XRCC4/XLF filaments: XLF/XRCC4 filaments have different curvatures with same and different XRCC4.XLF constructs.

DNA Ligase IV also interacts with XRCC4 in the helical tail between the residue 173 and 195. Interestingly, when DNA Ligase IV peptide is cocrystalised with XRCC4, the coiled-coil structure extends to cover the region of the helical tail (Sibanda *et al*., 2001). As for the C-terminal domain, it is likely to be dispensable for NHEJ activities (Mizuta *et al.*, 1997; Grawunder *et al.*, 1998; Modesti, Hesse and Gellert, 1999). Nevertheless, it contains the nuclear localisation signal (NLS) and an acidic cluster which promotes autotranscription (Mizuta *et al.*, 1997; Grawunder *et al.*, 1998). Also, the C-terminal domain of XRCC4 is the main domain of modification including the phosphorylation by DNA-PK mainly at S260 and S318 (Yu *et al.*, 2003; Lee *et al.*, 2004). In addition to DNA-PK phosphorylation, T233 on the CTD can be phosphorylated by CK2 and interacts with PNKP on the FHA domain (Koch *et al.*, 2004). S232 and T233 are also phosphorylated to interact with the FHA domain of APLF (Cherry *et al.*, 2015). This may apply to other proteins with FHA domain including Aprataxin (Clements *et al.*, 2004; Macrae *et al.*, 2008).

The second member of the superfamily, XLF, also known as Cernunnos, has 299 amino acids with a molecular weight of 33kDa and like XRCC4 can be divided into 3 domains: the head domain (1-135), the helical tail domain (136-233) and the C-terminal domain (234-299). Structural information is mostly limited to the head and helical tail domains (1-233). Like XRCC4, XLF also forms a homodimer through the helical tail domain to form the coiled coil, which also diverges in the middle. Unlike XRCC4, the helical tail domain of XLF is not just a single helix but three helices, of which the two C-terminal helices fold back to interact with itself. This also precludes DNA Ligase IV binding of XLF. The C-terminal domain is mostly unstructured but has the role of interacting with other proteins. For example, the C terminus region of XLF (281-299) interacts with Ku80 on the vWA domain to change the conformation of Ku70/80 (Nemoz *et al.*, 2018).

Interaction between XRCC4 and XLF takes place through hydrophobic interaction of the head domains. The XLF L115 docks in a hydrophobic pocket of XRCC4 formed by M59, M61, L108, and F106 (Hammel *et al.*, 2011). This head-to-head interaction is extendable and could result in a long filament and four independent groups showed similar left-handed XRCC4/XLF filament with a six-fold screw axe (Hammel *et al.*, 2011; Ropars *et al.*, 2011; Wu *et al.*, 2011; Andres *et al.*, 2012). However, the filaments have different curvatures, which are amplified

through the extended filament, due to the fact that the XLF docks on XRCC4 at different angles, indicating that the filament is elastic and flexible. Moreover, it was shown that the XRCC4/XLF filament could bind DNA (Yu *et al.*, 2003). There are hypotheses that the filament can form a positively-charged, elongated grooved channel that binds DNA (Yu *et al.*, 2003; Hammel *et al.*, 2011; Menon and Povirk, 2017). However, although the filament stands on its own without other NHEJ components *in vitro*, *in vivo* study showed that XLF was co-precipitated with not only XRCC4 but DNA Ligase IV (Ahnesorg, Smith and Jackson, 2006). In addition, the interaction of XLF and XRCC4 is reduced when DNA Ligase IV is absent. However, it remains unclear how DNA Ligase IV cooperates with the interaction of XLF and XRCC4 (Calsou *et al.*, 2003; Jayaram *et al.*, 2008). As for the disassembly of the filament, the phosphorylation of DNA-PKcs on XLF/XRCC4 can facilitate the process (Roy *et al.*, 2012a).

The last member of the XRCC4 superfamily, PAXX (Paralog of XRCC4 and XLF) was discovered in our group and characterised functionally in collaboration with the Jackson group (Ochi *et al.,* 2015). PAXX is the smallest member of this family and has 204 amino acid residues with a molecular weight of 22kDa. The head domain covers the region from residue 1 to 113; the helical tail domain is comprised of residues 114 and 145, and the rest (PAXX 146-204) is the flexible C-terminal domain. Structural information is available on region (1-142) before the unstructured C-terminal tail and PAXX shares the same structural organisation with XRCC4. However, unlike XRCC4 and XLF, PAXX does not interact with other members of the superfamily or form any high-order protein filament. Also, it has the preferred native state of homodimers in solution (Ochi *et al.*, 2015). PAXX interacts with Ku70/80 through the C-terminal domain and the key residues of V199 and F201 (Ochi *et al.*, 2015). Through this interaction, PAXX is involved to promote NHEJ (Ochi *et al.*, 2015). Later, it was shown that PAXX and XLF have redundant role in NHEJ and that the combined loss of XLF and PAXX can result in synthetic lethality in mammals (Balmus *et al.*, 2016; Kumar, Alt and Frock, 2016; Lescale *et al.*, 2016; Liu *et al.*, 2017). Moreover, in a side project of my PhD to understand the temporal organisation of NEHJ using single-molecule method, we found that PAXX actually participates in NHEJ at a very early stage and could help promote DNA end synapsis with Ku and DNA-PKcs (Wang *et al.*, 2018).

1.3.2.4 Artemis

Artemis belongs to the β-CASP family of the metallo-β-lactamase superfamily and SNM1 family of human proteins (de Villartay *et al.*, 2009). Artemis is a single peptide with 692 amino acid residues, which can be divided into two main parts—the N-terminal nuclease region (1-385) and the flexible C-terminal tail (385-692). The nuclease region is encoded by exons 1-13, while the whole C-terminal tail is encoded by exon 14 (Moshous *et al.*, 2001). The nuclease is composed of a metallo- β-lactamase domain, which contains the catalytic core of the nuclease (de Villartay *et al.*, 2009), and a β- CASP domain. Structures of β- CASP family proteins show that the β- CASP domain, inserted in the C terminal part of the metallo-β-lactamase domain acts as a cap to cover the catalytic core of the metallo-β-lactamase domain (Ochi, Wu and Blundell, 2014).

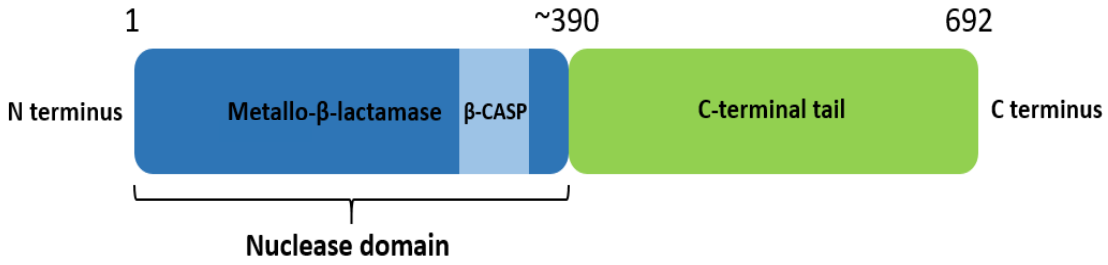


Figure 10. Schematic diagram of Artemis. Artemis has two domains-- the N-terminal nuclease domain and the C-terminal intrinsically disordered tail.

From the perspective of its physiological role, Artemis is the major nuclease involved in end-processing of NHEJ and Artemis-deficient cells show increased ionizing radiation sensitivity (Moscariello *et al.*, 2015). Moreover, Artemis is well known for being indispensable for V(D)J recombination, which is the fundamental immunological process creating diversity of antibodies and T cell receptors (Jung and Alt, 2004). Actually, it was first identified as the gene responsible for radiosensitive-severe combined immunodeficiency (RS-SCID) and Athabascan SCID (SCIDA) (Moshous *et al.*, 2001).

From the perspective of its enzymatic activity, Artemis has intrinsic 5′ exonuclease activity (Li *et al.*, 2014) and weak single-strand DNA (ssDNA) endonuclease activity, which can be greatly stimulated by interaction with DNA-PKcs (Goodarzi *et al.*, 2006). More importantly, once activated by DNA-PKcs, Artemis cuts the hairpin DNA produced by RAG1/RAG2 complex,

allowing the further ligation of coding joints. Furthermore, Artemis is the only vertebrate endonuclease reported so far that can open DNA hairpins (Chang and Lieber 2016).

There have been extensive studies on the nuclease activity of Artemis. It was shown that Artemis endonuclease activity varies with the nature of the substrate (Ma *et al.*, 2002). Artemis cuts at the intersection of dsDNA and ssDNA when the substrate is a 5′ overhang. When there is a 3′ overhang, Artemis cuts four nucleotides away from the intersection (Ma *et al.*, 2002). Moreover, Artemis can act on DNA hairpins and cut at a position on the 3′ side and two nucleotides away from the end (Ma *et al.*, 2002). Recently a new hypothesis that unites these various cutting patterns in one mechanism (Chang and Lieber 2016) has been proposed in which Artemis recognises three points of the substrate: points A, B and C. Point A sits on the 5′ to 3′ strand at the intersection of dsDNA and ssDNA. Point B is on the opposite side of point A. Point C, which is one nucleotide away from point B on the 5′ side, is the site where Artemis cuts. This hypothesis fits well with the biochemical analysis of Artemis based on different types of oligonucleotide digestion. However, the structure of neither the whole Artemis molecule nor the Artemis catalytic domain has been solved and it is unclear how Artemis works.

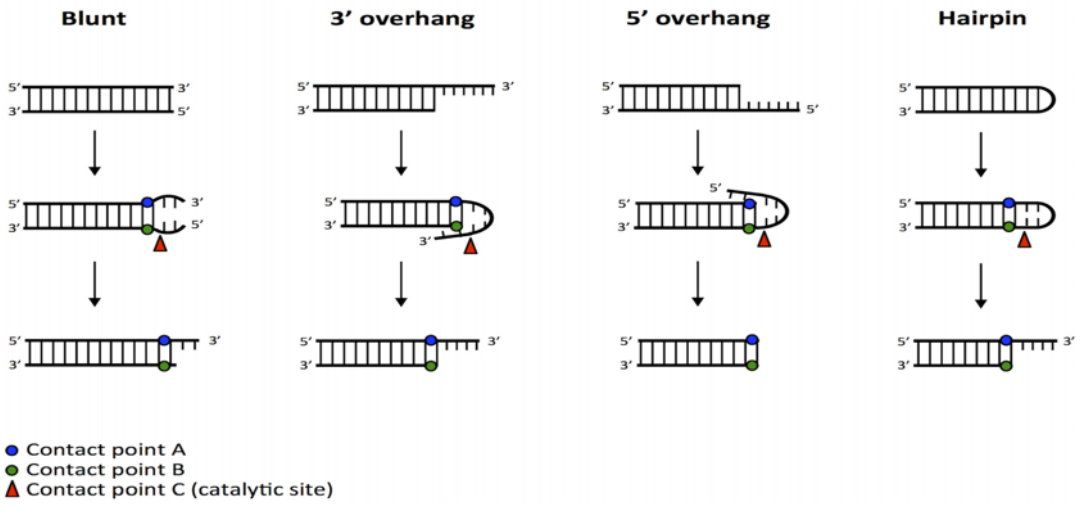


Figure 11. Hypothesis of Artemis endonuclease activity (Adapted from Figure 1 of Chang and Lieber 2016). To unite the cutting patterns of Artemis on different DNA ends, three points of the substrate are involved—point A, B and C. Point A (blue) is located at the intersection of dsDNA and ssDNA on the 5 ' to 3 ' strand. Point B (green) sits on the opposite site of point A. Next to point B on the 5 ' side, point C (red) is the catalytic site where Artemis cuts.

Comparison to homologues in the SNM1 family, including SNM1A and SNM1B, indicates that H33, H35 and H115 of Artemis probably coordinate a zinc ion. D37 and D136, possibly with E5 and E296, could be interacting with a Mg or Mn ion (Ma *et al.*, 2002; Pannicke *et al.*, 2004; Huang *et al.*, 2009). It has been shown that Artemis can function as an endonuclease without DNA-PKcs when Mn^{2+} ions are present *in vitro* (Chang and Lieber 2016). Nevertheless, it is uncertain which metal ions are involved in the enzyme activity.

Despite the large number of studies of the endonuclease activity of Artemis, the molecular mechanism of the endonuclease activity and how Artemis cuts the hairpin during V(D)J recombination still remain unclear 17 years after the discovery of the endonuclease-functional Artemis/DNA-PKcs complex (Ma *et al.*, 2002). The studies have focused on three aspects: protein-protein interactions within the system, phosphorylation of proteins involved, and crosstalk between these two aspects.

Early work showed that both the phosphorylation of Artemis and the physical presence of DNA-PKcs were important for the DNA-PKcs/Artemis complex to function as an endonuclease (Ma *et al.*, 2002). Later, the phosphorylation sites of DNA-PKcs on Artemis were examined by different groups and there were different results (Ma *et al.*, 2005; Goodarzi *et al.*, 2006; Soubeyrand *et al.*, 2006). The first group showed that there were 11 DNA-PKcs-phosphorylation sites on Artemis in addition to the three basal phosphorylation sites on Artemis defined earlier (Ma *et al.*, 2005). Many of these phosphorylation sites do not belong to the SQ/TQ cluster domains (SCD), which were predicted to be the phosphorylated sites of PIKK (Ma *et al.*, 2005). However, the work of another group later showed something different (Soubeyrand *et al.*, 2006). *In vitro* study showed that DNA-PKcs can phosphorylate on multiple sites while *in vivo* study showed that the main phosphorylation sites on Artemis had a different profile (Soubeyrand *et al.*, 2006). Also, it was shown that the region of Artemis 399-404 is important for the physical interaction between Artemis and DNA-PKcs (Soubeyrand *et al.*, 2006). Especially, residues L401 and R402 are indispensable to the interaction as L401G/R402N double mutant was not able to pull down DNA-PKcs from the whole cell extract (Soubeyrand *et al.*, 2006).

However, the region of Artemis interacting with DNA-PKcs was not defined and there was no information about which domain of DNA-PKcs Artemis interacts with. Strikingly, it was also shown that the formation of DNA-PKcs/Artemis complex and phosphorylation on certain Artemis residues are dispensable for the endonuclease function of Artemis (Soubeyrand *et al.*, 2006). Meanwhile, a third group had similar results on the study of DNA-PKcs and Artemis (Goodarzi *et al.*, 2006). It had similar phosphorylation profiles *in vitro/ in vivo* compared to the previous study (Soubeyrand *et al.*, 2006) but different to the early one (Ma *et al.*, 2005). It was proposed that the differences between these studies may be due to the fact that a lower salt concentration was used in the early experiments (Ma *et al.*, 2005) and that may cause the differences that are not physiologically sensible (Goodarzi *et al.*, 2006). Furthermore, *in vitro* experiments showed that the endonuclease activity of Artemis is supported by Ku, DNA-PKcs and ATP but not ATM as another group earlier showed that Artemis is also phosphorylated by ATM (Chen *et al.*, 2005; Goodarzi *et al.*, 2006). Not withstanding the differences, there was one thing in common from all these independent studies that phosphorylation on the Artemis C-terminal tail by DNA-PKcs (or other PIKKs including ATM) was not likely to have an effect on its endonuclease activity (Ma *et al.*, 2005; Goodarzi *et al.*, 2006; Soubeyrand *et al.*, 2006). It was proposed at that time that the physical presence of DNA-PKcs and/or its autophosphorylation are more relevant to the nuclease activity of Artemis (Ma *et al.*, 2005; Goodarzi *et al.*, 2006; Soubeyrand *et al.*, 2006). Six years later, it was shown that the C-terminal domain of Artemis is actually important for the hairpin cleavage in V(D)J recombination through its interaction with both DNA-PKcs and DNA Ligase IV (Malu *et al.*, 2012). In Artemis-deficient human pre-B cells, Artemis W489A, which loses interaction with DNA Ligase IV, and Artemis L401G/R402N, which loses interaction with DNA-PKcs, showed reproducible but moderate decreases in the coding joint formation (Malu *et al.*, 2012). However, Artemis W489A/L401G/R402N triple mutant had a strong inhibition of the formation of a coding joint (Malu *et al.*, 2012). This is in agreement with the previous studies and it seems that DNA-PKcs and DNA Ligase IV are playing redundant roles to recruit Artemis to the DNA end. Moreover, it was recently shown *in vivo* that the kinase activity of DNA-PKcs is not required for hairpin opening (Jiang *et al.*, 2015). Instead, it is the phosphorylation of ATM on DNA-PKcs (or Artemis) that helps recruiting Artemis to the DNA end for processing. In addition to the previous aspects mentioned, attention was paid to the role of Ku for the endonuclease function of Artemis as Ku is the most important binding partner of DNA-PKcs.

There are studies concerning the effect of Ku on Artemis. It was proposed that Ku is important for Artemis endonuclease activity as it activates DNA-PKcs for autophosphorylation (Goodarzi *et al.*,2006). It was also proposed that Ku cannot bind to DNA-PKcs/Artemis complex without DNA ends or when the kinase activity is inhibited, but the DNA/Ku/DNA-PKcs/Artemis complex has a specific phosphorylation profile (Drouet *et al.*, 2006). Moreover, it has been shown that Ku 80 C terminus plays an important role in Artemis-mediated processing of DSB ends via DNA-PKcs, although it does not come into contact with Artemis (Weterings *et al.*, 2009). However, further details of the mechanism remain confusing as the latest result showed that the autophosphorylation of DNA-PKcs is not important in Artemis recruiting or processing (Jiang *et al.*, 2015). The crosstalk of phosphorylation and protein-protein interaction becomes more and more complicated with increasing numbers of the NHEJ members involved or post-tanslational modification involved, including DNA-PKcs, Artemis, Ku, DNA Ligase IV and ATM. Nevertheless, how they affect Artemis function as an endonuclease still remains unknown. This is not only due to the complexity of Artemis and its interaction network, but also due to the fact that the details of other relevant components and their internal interactions remain unclear. Therefore, to understand how Artemis functions as an endonuclease, it is important to understand two things: the structure of the nuclease domain of Artemis and the interaction between Artemis and DNA-PKcs, the most important partner and activator of endonuclease activity.

Yet this is not the end of the story of Artemis. Artemis is interacting with other proteins through the C-terminal flexible domain besides DNA-PKcs. For example, as mentioned, Artemis 485-495 is important for the interaction between Artemis and the DNA binding domain of DNA Ligase IV (Malu *et al.*, 2012). Later, Ochi, Gu and Blundell (2013) showed that Artemis residues 485-495 undergo concerted folding and binding with the first two helices of DNA Ligase IV to form a three-helical bundle. When interacting with DNA Ligase IV, W489 of Artemis forms a hydrogen bond with D18 of DNA Ligase IV, stabilized by interaction between F492/F493 of Artemis and F49/F42 of DNA Ligase IV. Moreover, P487 of Artemis sits in the hydrophobic pocket formed by L53, A52 and F49 of DNA Ligase IV (Ochi, Gu, and Blundell 2013). There are also other binding partners with no structural information. For example, residues 641-660 of Artemis can bind to the second BRCT domain of an adaptor protein Pax transcription-activation-domain interacting protein (PTIP) via phosphorylation (Wang *et al.*,

2014). Furthermore, Artemis physically binds to the Mre11/Rad50/Nbs1 complex (MRN complex) in an ATM-dependent way under Ionizing-radiation induction (Chen *et al.*, 2005). Although the region is not yet defined, it is likely that the C terminus will be the interacting part given that hyperphosphorylation of Artemis is associated with Artemis-MRN complex interaction (Chen *et al.*, 2005).

These observations relate to the interesting phenomenon of intrinsically disordered regions, which are observed in many proteins participating in regulatory processes. Intrinsically disordered regions have not been as well investigated as folded proteins in the past despite the fact that they are common and many intrinsically disordered regions are highly conserved (Dyson and Wright, 2005). This is surprising given that Dunker and colleagues in 2002 (Dunker *et al.*, 2002) pointed out that functions of intrinsically disordered polypeptides include regulation of transcription and translation, cellular signalling and regulation of large-multiprotein-complex assembly.

The intrinsically disordered C-terminal region of Artemis recruits other factors and facilitates the assembly of the multi-component NHEJ complex. It appears to act like a string “tying” together these other components with the covalently linked N-terminal nuclease region. Besides Artemis, other examples include APLF, which also has disordered regions interacting with other components of NHEJ including Ku (Grundy *et al.*, 2012), are also likely to act as a flexible string that hold the factors closed to the NHEJ site.

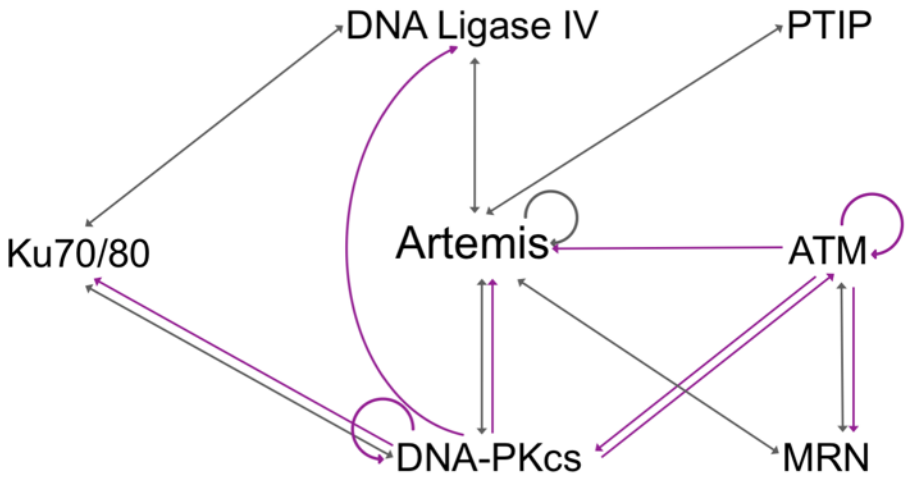


Figure 12. Interaction network of Artemis. The double arrows coloured ink indicate protein-protein interaction and the single arrows coloured purple indicate phosphorylation modification.

1.3.2.5 DNA Ligase IV

Human DNA Ligase IV is an ATP-dependent DNA ligase, belonging to the nucleotidyltransferase superfamily like all the human DNA ligases (Shuman and Lima, 2004; Alan E. Tomkinson *et al.*, 2006; Ellenberger and Tomkinson, 2008). Compared to other human DNA ligases including DNA Ligase I and DNA Ligase III, it has some unique features. DNA Ligase IV is the only human DNA ligase involved in NHEJ as a core NHEJ component and it is dedicated to NHEJ. It performs DNA ligation in a three-step process of nucleotidyl-transfer reaction- DNA Ligase IV adenylation, adenyl transfer to DNA and DNA backbone sealing (Alan E. Tomkinson *et al.*, 2006; Shuman, 2009). In the first step of DNA Ligase IV adenylation, ATP interacts with a lysine residue of DNA Ligase IV to form the lysine-adenosine monophosphate (AMP) covalent intermediate. Then in the second step, the active lysine-AMP intermediate attacks the 5′ phosphate on the DNA end and transfers the AMP to the 5′ end. In the last step, the 3′ hydroxyl group of the DNA reacts with the 5′ phosphate-AMP intermediate to rebuild the phosphodiester bond of the broken DNA and release the AMP (Alan E. Tomkinson *et al.*, 2006; Shuman, 2009). It was recently shown that NAD+ could also stimulate the adenylation of DNA Ligase IV for further DNA ligation (Chen and Yu, 2019).

DNA Ligase IV comprises five domains: DNA binding domain (residues 6-239) (DBD); Nucleotidyltransferase domain (residues 240-453) (NTase); OB-fold domain (residue 458-606) (OBD); and two BRCT domains (residues 654-740; residues 815-911) linked via a flexible region (residues 741-814) (Figure 13). Together, DBD, NTase and OBD are considered as the catalytic core of DNA Ligase IV. This catalytic core is highly conserved among the human DNA ligases including DNA Ligase I, DNA Ligase III and DNA Ligase IV (Ellenberger and Tomkinson 2008).

Starting from the N terminus of DNA Ligase IV, DBD is formed of 12 α helices. As the name implies, DBD binds DNA non-specifically (Pascal *et al.*, 2004). It shares a similar structure with other human DNA ligases but has a relatively long helix α2 and specific conserved residues to interact with Artemis (De Ioannes *et al.*, 2012). The first two helices interact with Artemis C-terminal region to form a three-helical bundle to strengthen the interaction of Artemis on DNA ends (De Ioannes *et al.*, 2012; Malu *et al.*, 2012). The NTase is a common domain among

all DNA and RNA ligases with a similar structure (Shuman and Lima, 2004; Ochi, Gu and Blundell, 2013; Kaminski *et al.*, 2018). It comprises two parts: one binds to the 3′ end of the broken DNA and the other binds to the 5′ end. The lysine (Lys 273) that is adenylated for further DNA ligation is located at the centre of the NTase domain to target the DNA breakage site (Ochi, Gu and Blundell, 2013; Kaminski *et al.*, 2018). OBD is also conserved among all DNA ligases. It contains a β-barrel domain with five anti-parallel β-strands and important for forming the lysine-AMP intermediate for all DNA ligases (Ochi, Gu and Blundell, 2013; Kaminski *et al.*, 2018).

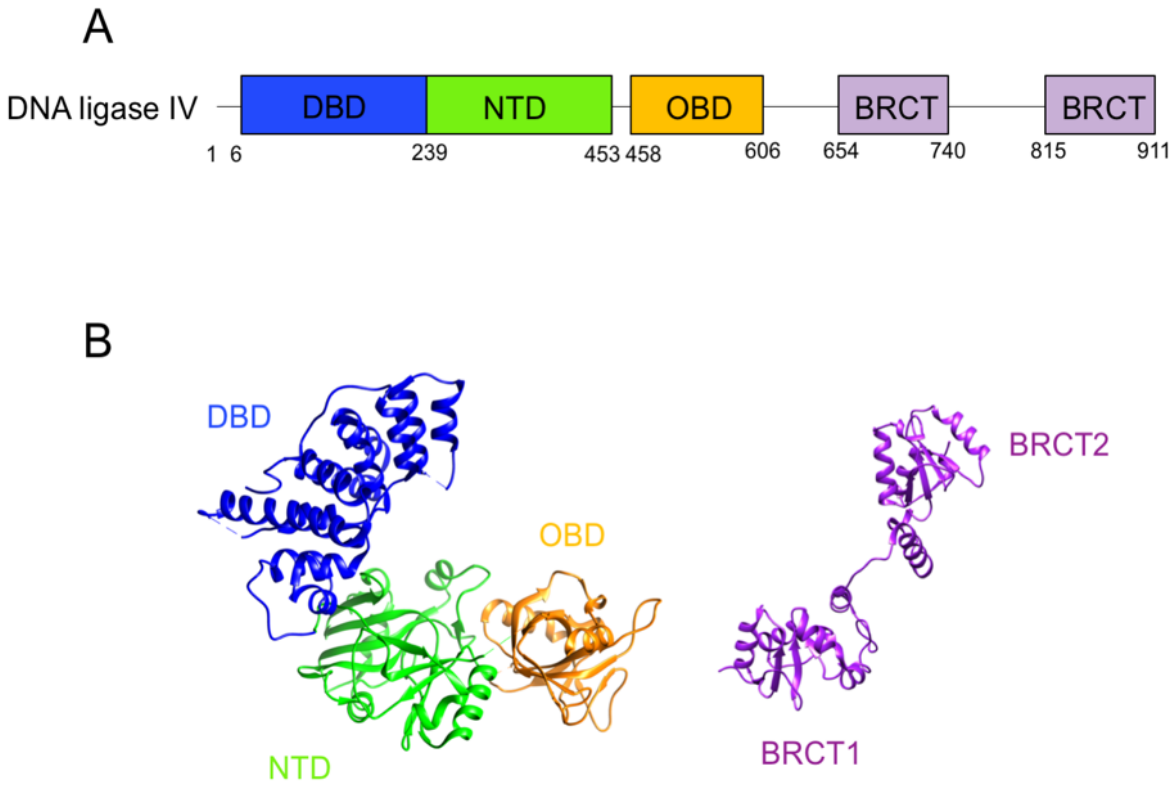


Figure 13. **Structural information of DNA ligase IV**. (A) Schematic diagram of DNA ligase IV. The DNA ligase IV is formed of the catalytic core (1-606) and C-terminal region (607-911). The core contains three domains-- DNA binding domain (DBD), Nucleotidyltransferase domain (NTD) and the OB fold domain (OBD). The C-terminal region contains two BRCT domain linked by a flexible linker; (B) Structures of the catalytic core (PDB: 3W5O) and the BRCT domains without the stabilising XRCC4 (PDB:3II6) of DNA ligase IV (Ochi et al., 2013; Wu et al., 2009).

So far, there are four structures of the catalytic domain in different conformations: apo-DNA Ligase IV catalytic domain (residues 1-609) (PDB: 3W5O); DNA Ligase IV catalytic domain (residues 1-609) in complex with Artemis peptide (residues: 485-495) (PDB: 3W1G); DNA Ligase IV catalytic domain lysyl-adenylate (residues 1-620) in complex with nicked DNA (PDB: 6BKF) and DNA Ligase IV catalytic domain (residues 1-620) in complex with DNA-adenylate (PDB: 6BKG) (Ochi, Gu and Blundell, 2013; Kaminski *et al.*, 2018). In the apo-form (PDB: 3W5O), the catalytic region is in an "open" state which is stabilised probably by the interaction area of 1,104 $Å^2$ between NTase and OBD (Ochi, Gu and Blundell, 2013). Binding of the Artemis peptide does not change the conformation, indicating that this protein-protein interaction may not stimulate DNA ligation (Ochi, Gu and Blundell, 2013). In the structure of Ligase IV catalytic domain lysyl-adenylate (residues 1-620) in complex with nicked DNA, the catalytic domain remains in the "open" status (Kaminski *et al.*, 2018). At this stage, DNA Ligase IV recognises the DNA nick through the interaction between Arg293/Arg443/Lys449/Lys451 and the 5′ phosphate and the following backbone phosphate downstream (Kaminski *et al.*, 2018). While the 5′ phosphate sits next to the catalytic site, the closest 5′ phosphate oxygen is around 5.4 Å away from the AMP phosphorus atom, which is still far for the nucleophilic attack and adenyl-transfer from Lys273 to DNA (Kaminski *et al.*, 2018). This indicates that structural repositioning of the Ligase IV lysyl-AMP or DNA 5′ phosphate is required of the further ligation steps. Actually, in the structure of DNA Ligase IV catalytic domain (residues 1-620) in complex with DNA-adenylate, the catalytic domain then adopts a "closed" conformation around the nicked DNA (Kaminski *et al.*, 2018). Alignment of the two structures of Ligase IV 1-620 in complex with nicked DNA showed that the conformation of the DBD and NTase subdomains mostly remain the same while the OBD swivels to encircle the DNA substrate completely and to interact with DBD, forming a "latch". However, it remains unclear what triggers the conformational change of the catalytic domain from "open" to "closed".

Next to the catalytic domain is the C-terminal domain containing two BRCT domains. BRCT domain contains four parallel β sheets flanked by one α helix and two α helices on either side (Bork *et al.*, 1997). BRCT repeats are common and frequently found in DDR proteins including 53BP1, BRCA1, BARD1, PARP-1 and TopBP1 which were mentioned previously (Williams, Green and Glover, 2001; Derbyshire *et al.*, 2002; Joo, 2002; Birrane *et al.*, 2007; Loeffler *et al.*,

2011; Rappas, Oliver and Pearl, 2011). The tandem BRCT repeats are known to bind phospho-peptides while it is unclear how phospho-peptides interacts with the BRCT repeats of DNA Ligase IV though they were expected to interact with Ku70/80 complex (Manke, 2003; Rodriguez *et al.*, 2003). Moreover, the tandem BRCT repeats are known to function as protein-protein interaction domains and the linker between one BRCT domain and the second is also important for protein-protein interaction (Watts and Brissett, 2010). A good example here is the interaction between DNA Ligase IV and XRCC4. The first structure of the DNA Ligase IV/XRCC4 complex revealed the core XRCC4 interaction domain (XID) is the linker between BRCT1 and BRCT2 (DNA Ligase IV 748-784) (Sibanda *et al.*, 2001). The binding of DNA Ligase IV interacts on XRCC4 changes the conformation of the coiled coil but what effect it has remains unclear (Sibanda *et al.,* 2001). Later, a structure with longer construct of DNA Ligase IV C-terminal region showed that, instead of just the XID, BRCT2 is also interacting with XRCC4 and this interaction is necessary for the stabilisation of DNA Ligase IV/XRCC4 complex in cell (Wu *et al.*, 2009). In addition to the influence on protein-protein interaction, the BRCT1 domain is also proposed to be important for NAD+ recognition and adenylation of DNA Ligase IV while no structure is available (Kaminski *et al.*, 2018).

1.4 Structure in Biology

1.4.1 Structure, Function and Drug Discovery

Studying structure provides detailed knowledge and insights that are not only important for understanding the structure and function of important biological molecules but are also central to selection of targets for structure-guided drug discovery. For example, understanding the structure of a protein and its evolution can provide us with the detailed mechanism of how the protein fulfils its physiological roles. For example, the catalytic sites of the nucleases SNM1A and SNM1B are highly structurally conserved, indicating a similar mode of exonuclease catalysis. Furthermore, we can usually locate the functional site of a protein based on the structure and evolutionary conservation. For example, for DNA-PKcs, although there was no structural information of ATP or substrate binding, based on the available information of the existing structures of other PI3-kinases, we can have an idea of the mode of phosphorylation of DNA-PKcs.

The connection and interplay between structure and function are not just limited to catalytic pockets and enzyme activity. Many proteins have non-enzymatic functions which are also highly connected to the structures. A good example would be the discovery of PAXX, which was discovered in our group (Ochi *et al.*, 2015). Although there was no known biological role of PAXX, it was predicted to be involved in NHEJ due to the structural similarity with other XRCC4 superfamily members. Later collaboration on cellular study showed that PAXX interacts with Ku and is involved in NHEJ. Therefore, understanding the structure can be significant for predicting the function of the protein with unknown physiological role, guiding the subsequent cellular physiological study.

Furthermore, structure on its own can provide us with indications of potential roles of the protein. In the case of DNA-PKcs, the kinase domain is only around 300 residues at the C-terminus of a protein comprising 4128 residues. The structure demonstrated that it sits on the head unit of the whole molecule, with other domains including N-terminal bridge and circular cradle acting as stages on which other molecules assemble. Recently it has been shown that the N-terminal arm is involved in the allosteric regulation of kinase activity.

However, there are still structural domains of DNA-PKcs with unknown functions, indicating that other functions of DNA-PKcs may not have been revealed.

Biological function can also be mediated by intrinsically disordered regions, which are common cell regulatory systems and often are highly conserved (Dyson and Wright, 2005). Dunker and colleagues in 2002 (Dunker *et al.*, 2002) pointed out that functions of intrinsically disordered polypeptides include regulation of transcription and translation, cellular signalling and regulation of large-multiprotein-complex assembly. In NHEJ, there are many intrinsically disordered regions that appear to play the regulatory roles, some providing access to short foldable regions and others linking different components in a flexible manner. These include the XLF domain interacting with Ku70/80, APLF domain interacting with Ku70/80 and Artemis domain interacting with DNA Ligase IV.

Studying the structure can also be very important for understanding diseases and drug discovery. Folding of a protein allows residues that are distant in primary structure to interact with each other; this is difficult to understand without 3D structure determination. With the structural information of the driver protein of a specific disease, the mutations from patients can be located, producing a clear picture of the molecular origins of the disease. Moreover, the structure of target proteins with a drug candidate or antibody could provide a detailed mode of action (MOA). Furthermore, structures of the protein can be used for high-throughput screening (HTS) and fragment-based drug discovery (FBDD).

In fact, from the perspective of oncology, the NHEJ system is a good target. Biologically, one of the underlying hallmarks of cancers is the genomic instability, which leads to a higher propensity to accumulate DNA damage (O'Connor, 2015). To make use of this hallmark, DNA-damaging radiotherapy and chemotherapy were introduced some time ago to induce further instability in cancer cells and to kill them. However, in various cancers, the expression of DSB repair genes is altered. Activation and upregulation of DSB repair genes is one of the reasons for the resistance of radiotherapy and chemotherapy (Srivastava and Raghavan, 2015). Considering that NHEJ is the predominant DSB repair pathway in the mammalian system, targeting NHEJ would be attractive strategy to treat cancer especially in combination with radiotherapy or chemotherapy. Many pharmaceutical companies are targeting DNA repair for

cancer treatment including Astra Zeneca and Merck, and one key target is DNA-PK of NHEJ (https://www.astrazeneca.com/what-science-can-do/stories/ddr.html). In addition to DNA-PK and other enzymes in NHEJ, the protein-protein interaction sites in NHEJ may be also good starting points for drug discovery. Structurally, the interactions between globular domains and intrinsically disordered regions tend to be the most druggable protein-protein interactions (Jubb, Blundell and Ascher, 2015). In this respect, the interaction between Artemis and DNA Ligase IV is of great interest. The fact that DNA Ligase IV is the only ligase that joins DSB ends in NHEJ makes it a potential candidate for therapeutics that target NHEJ specifically.

1.4.2 Methods of Structure Determination

To solve the structure of protein and to build the final atomic model, three different methods are most often used: X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, and electron microscopy (EM). These methods present different challenges and each has their own forte.

In the case of X-ray crystallography, the protein requires purification, crystallisation and stability in an intense beam of X-rays. As protein molecules are packed symmetrically and repeatedly in crystal, they will produce an X-ray diffraction pattern with the amplitude of X-ray wave recorded. Together with other methods to determine the phase, the diffraction pattern must then be analysed to determine the electron distribution within the molecule to obtain the final result of the map of electron density. The map can then be interpreted to locate the position of each atom. X-ray crystallography is excellent at providing atomic information with fine details showing every single atom in the molecule. However, the protein has to be crystallised for X-ray crystallography and this can be difficult or even impossible in many cases. It can take years to find the ideal condition, if there is any, for the protein to be crystallised. Classic examples include the ribosome and DNA-PKcs in NHEJ which took over a decade to produce decent crystals. In general, X-ray crystallography is the method to study the structure of relatively stable and ordered proteins that form nice crystals. It has also been the method for routine structure-based drug discovery.

Instead of showing the exact location of each atom, NMR spectroscopy gives information about the local conformation and the distance between atoms that are close to one another. For NMR experiments, the protein is first purified, then placed in a strong magnetic field and probed with radio waves. A characteristic set of observed resonances can be analysed to give the list of atomic nuclei close to one another and to specify the local conformation of atoms bonded together. The obtained restraints can then be used to build the model of the protein showing the relative location of each atom. NMR has the advantage that it provides information of protein in solution and it is the preferred method to study the structure of flexible protein. However, NMR is limited by the size of the protein. It is not useful for large

proteins due to the sample behaviour in solution and the problem of signal peaks overlapping in the spectra.

The most often used EM is either negative staining or cryo-EM, based on whether the sample is treated with negative stain. High-resolution structures can be defined using cryo-EM, which is the rapidly developing technique of structure study. Cryo-EM is the main technique used for structure study in this research and more details will be introduced in the section 1.4.3. Unlike X-ray crystallography and NMR, EM images the protein directly using electron beams. For EM, the protein is purified, loaded on grids and placed under electron microscopes. When the grid is placed in microscope, there are two ways of collecting data for subsequent analysis: single particle analysis (SPA) and cryo-electron tomography (cryo-ET). Currently, protein structures by EM at atomic resolution are mostly done using single-particle analysis. To obtain high resolution in single-particle analysis, a large number of protein particles, usually tens of thousands to millions, are extracted from the micrographs to get clear 2D projections from all different angles, which will later be used to reconstruct the 3D model. Cryo-EM has the advantage that it does not require protein crystals and can visually examine the protein particles directly. Furthermore, cryo-EM allows the protein to stay in native condition without restraints such as crystal packing and can tolerate flexibility of the protein to a certain extent. However, it still remains highly challenging to achieve high resolution for flexible regions of the protein. Moreover, cryo-EM is currently not good at solving structures of small protein particles as it will be difficult to identify the angles of projection of small particles for 2D and 3D classification. Nevertheless, many developments are undergoing in the field of cryo-EM and it has shown the potential of solving structures of small proteins including the record-keeping 52 kDa streptavidin (Fan *et al.*, 2019).

In addition to the three methods for structure determination, there are many other facilitating methods that are not as definitive but could facilitate the structure study by adding restraints and providing biophysical and biochemical information, including Circular Dichroism (CD), Analytical UltraCentrifugation (AUC), and a series of different Mass Spectrometry (MS) approaches, including native MS, crosslinking MS, and Hydrogen–Deuterium eXchange (HDX). The details of these techniques will not be discussed but it is important to point out that all those techniques together can provide a powerful toolbox for

structure study of various types of proteins and complexes and a new integrated structural biology is on the horizon.

1.4.3 Cryo-EM Resolution Revolution

Cryo-EM is the main method I have been using in my PhD to study the structure of NHEJ complexes and it has been developing at a very high speed. The breakthrough of cryo-EM to become a routine method for solving structures at atomic resolution came around 2013 when the direct electron detector was launched. After that, a series of structures including TRPV1 ion channel, F420-reducing [NiFe] hydrogenase and the large subunit of the yeast mitochondrial ribosome were solved at resolutions of 3.2- 3.4 Å, heralding the beginning of the new era of cryo-EM and structural biology (Li *et al.*, 2013; Liao *et al.*, 2013; Allegretti *et al.*, 2014; Amunts *et al.*, 2014). Compared to x-ray crystallography and NMR spectroscopy, cryo-EM is a new member of the "near-atomic resolution" club.

In general, the cryo-EM workflow can be separated into three steps including sample preparation, data collection and data analysis (Figure 14). In sample preparation, the aqueous sample is first loaded on specific cryo-EM grids. The grid is comprised of two components, the metal grid bar and a layer of covering membrane with empty holes. There are three common parameters to take into consideration including the metals of the grid bar, the material of the covering membrane and the pattern/size of the empty holes on the membrane. Common metals of the grid bar include copper and gold. As for the covering membrane, usual options are carbon (e.g. Quantifoil) and gold (e.g. Ultrafoil). The hole shapes and arrays can vary from precise and well-defined (e.g. 1.2/1.3) to open and disordered (e.g. Lacey carbon). In addition, there are many other advanced parameters of grid types including surface modification like PEGylation and extra supporting films such as graphene oxide (GO). Different proteins behave differently on the grids and there is no one grid that suits all samples. Therefore, the parameters mentioned should be considered for optimisation for sample preparation.

The aqueous sample (usually 2µl-3µl) is first applied to the grid and blotted with filter paper to remove the excess of liquid under a specified force of blotting for a defined period of time with a machine (e.g. Vitrobot). The grid is then plunge-frozen in liquid ethane to vitrify the aqueous sample which can then be observed using cryo-EM.

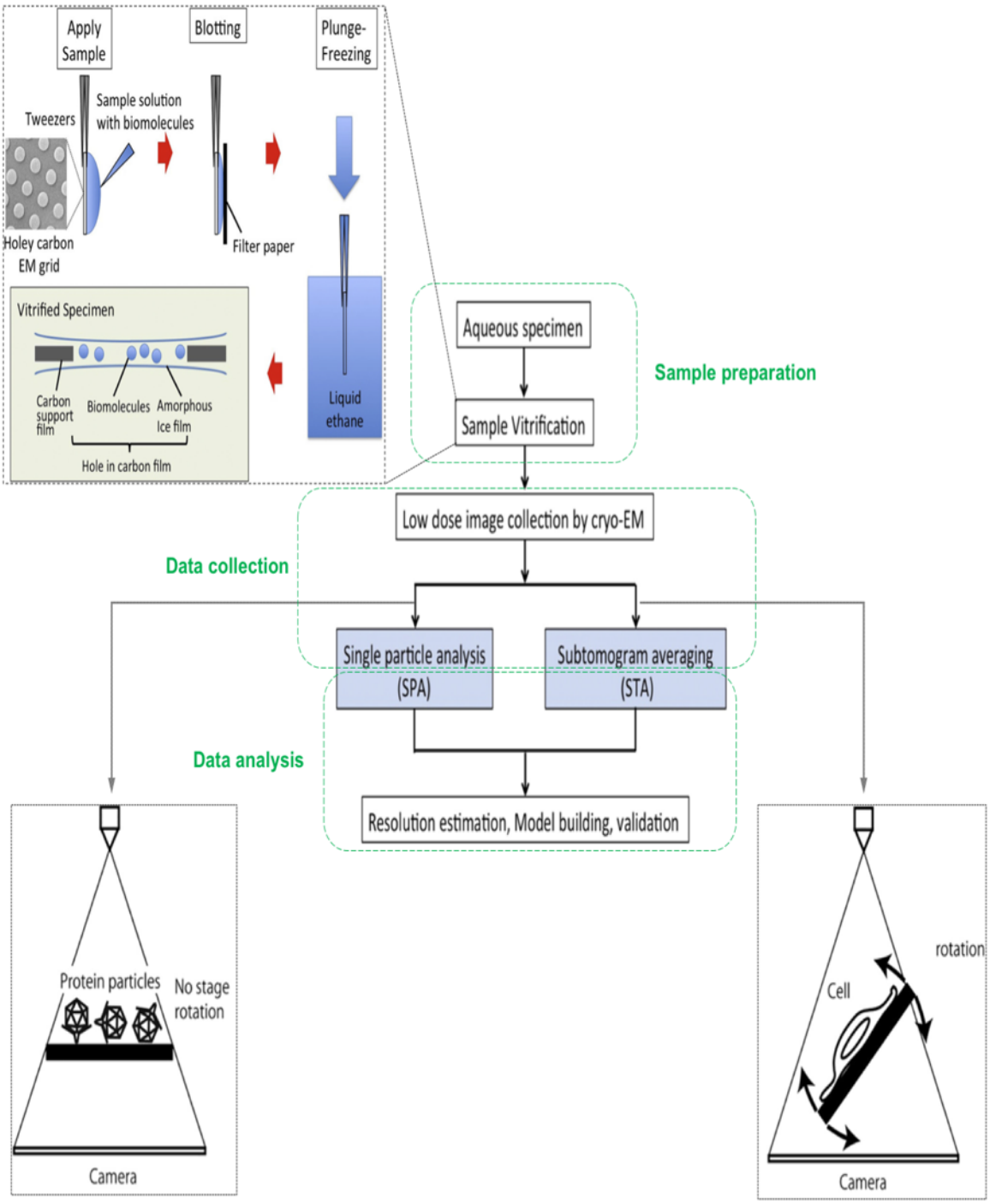


Figure 14. Cryo-EM workflow (Adapted from Figure 2 of Murata and Wolf, 2018). The three steps of cryo-EM include sample preparation (aqueous sample and sample vitrification), data collection (low dose image collection by cryo-EM for single-particle analysis or subtomogram averaging) and data processing, which will be introduced in detail later.

After sample preparation comes the data collection. Cryo-EM is similar to the EM used in the field of Material Science or Engineering as the same technique of transmission electron microscope (TEM) but running under the cryo-environment using liquid nitrogen. To achieve an electron beam with high energy and short wavelength, high accelerating voltages are used in cryo-EM, for example 200kV (e.g. Thermo Fisher Scientific Talos Arctica/ Glacios; JEOL CRYO ARM 200) and 300kV (e.g. Thermo Fisher Scientific Titan Krios; JEOL CRYO ARM 300). However, unlike many samples from Material Science or Engineering, the biological samples are very fragile, with low tolerance to the radiation damage caused by the electron beam. Therefore, cryo-EM uses a low-dose image collection strategy. In addition, there are two ways of data collection with stage rotation or without stage rotation.

Stage rotation is involved in cryo-electron tomography (cryo-ET). In this case, the grid is tilted for a series of angles (e.g. From -60 ° to +60 °) to record a stack of images known as the 'tiltseries'. The tiltseries is then aligned to the common rotation axis to build the 3D tomogram via weighted back-projections or other methods. The sub-tomograms of the macromolecules can then be extracted from the built tomogram for alignment and averaging to reconstruct the map of the macromolecules at a finer level.

No stage rotation is used in single particle analysis (SPA). In this case, instead of collecting tiltseries, movies of subframe collection are recorded with different defocus values for subsequent analysis. Compared to cryo-ET, the resolution revolution was mainly driven by the improvement of all aspects of SPA, which defined cryo-EM structures at near atomic level regularly. SPA is also the method used for cryo-EM experiment in my PhD project and will be introduced in detail in the following paragraphs (Figure 15).

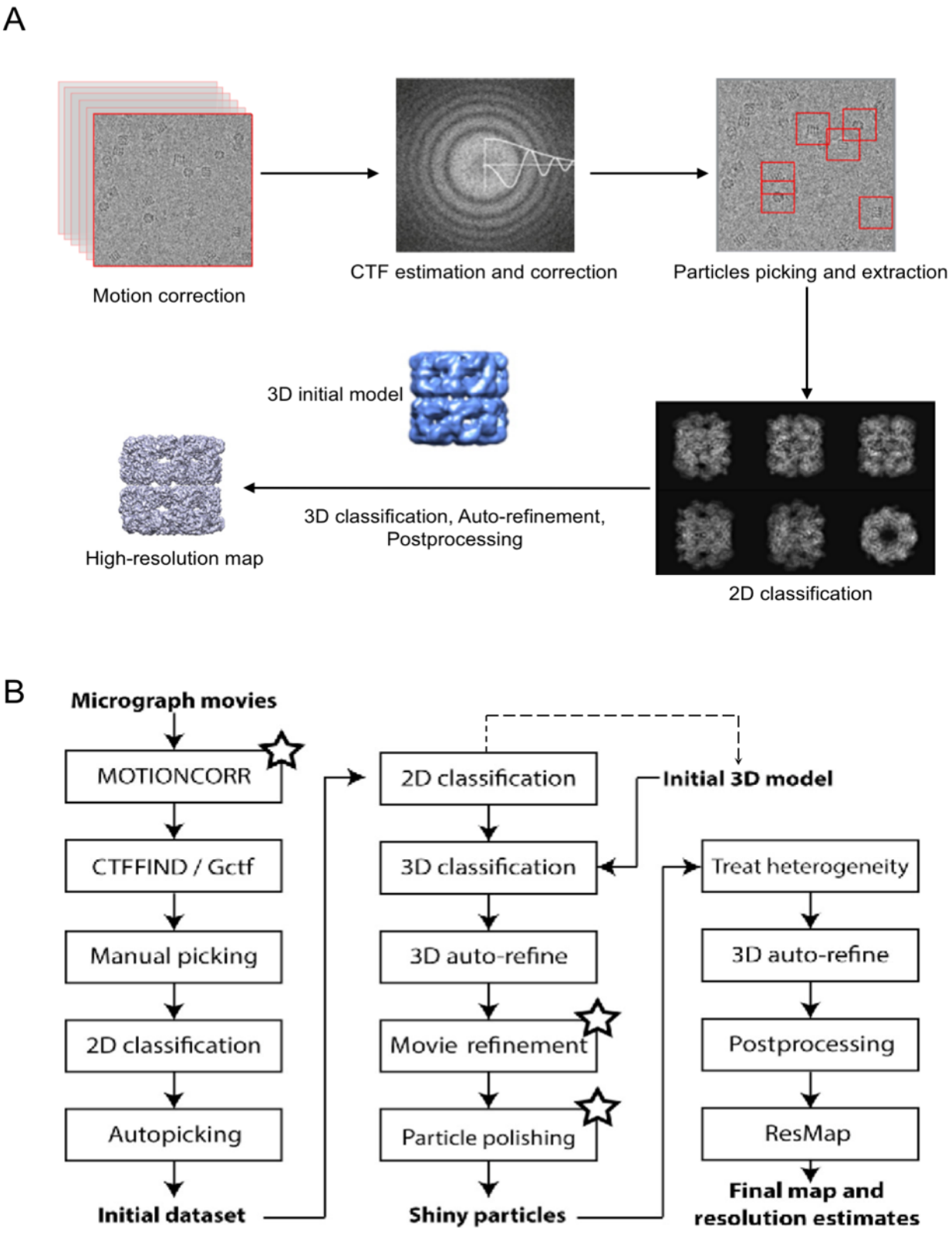


Figure 15. The processing of SPA. (A) Schematic diagram of the main steps and elements of SPA from the original movies to the final high-resolution map (Modified from Figure 2 of Carroni and Saibil, 2016). (B) Workflow of Relion 3.0 to achieve high-resolution cryo-EM map using SPA. The star labels emphasize on the new function of Relion 3.0 (Modified from Figure 1 of Scheres, 2016).

In general, the data processing of SPA is flexible and varies depending on the behaviour of the sample but can be separated into two parts: micrographs processing and particle processing. Micrographs need aligning, averaging and correcting first, before being used for particle picking and 2D, 3D analyses (Figure 15A).

With the rapid development of cryo-EM, there are many software packages now on the market for SPA and common ones include Relion, eMAN, Scipion, Sphire, cryoSPARC, cis-TEM and SIMPLE (Tang *et al.*, 2007; Elmlund and Elmlund, 2012; Scheres, 2012; de la Rosa-Trevín *et al.*, 2016; Moriya *et al.*, 2017; Punjani *et al.*, 2017; Grant, Rohou and Grigorieff, 2018). Different packages have their own advantages on data processing and different hardware requirement. For example, cis-TEM and SIMPLE do not require GPU for the data processing. During my PhD project, Relion has been the major package used for data analysis and it is one of the dominant software in the field (Figure15 B).

To do SPA in Relion, micrographs are first processed with motion correction and contrast transfer function (CTF) estimation and correction. All the movies of the subframes are traced for the beam-induced movement, aligned and averaged to produce the motion-corrected micrographs. The micrographs are then sent for CTF estimation and correction. All the micrographs with good CTFs will be selected.

After CTF correction, a subset of micrographs with good CTFs are chosen. The user should pick the particles of interest from them using a self-defined mask and later extract them with a self-defined box. These particles will be then applied to 2D classification to generate 2D projections of the target protein. This could give a fast feedback of the quality of the dataset and also produce templates for the following auto picking on all micrographs. In auto picking of the full dataset, hundreds of thousands of the particles can then be picked and again be subjected to 2D classification to obtain 2D projections of the full set of particles and clean up the contamination within the particles picked.

A subset of the particles with good 2D projections will be selected for 3D processing with an 3D initial model. The optimal model can be generated from previously existing structure models (e.g. from X-ray crystallography) or built from the previous 2D classes. The 3D

classification can then be done to distinguish the heterogeneity and conformational changes within the chosen particles to identify subsets that have different 3D conformations or include contamination. The nice subsets would be further subjected to high-resolution 3D auto-refine for details and then sent to postprocessing for sharpening. The particle processing does not have to stick to the classic workflow and could be done iteratively to optimise the subsets of particles and the 3D maps.

After obtaining the optimised final map, the final step is modelling and validation. Currently, there are some cryo-EM modelling packages that are transformed X-ray crystallography modelling packages, including CCP-EM and Phenix. Although the property of the final maps of cryo-EM and crystallography are different, the modelling stage is mainly similar and will not be described in detail.

1.5 Project Objectives

The objectives of my research are to study the biochemistry and structure of the human DNA-PKcs/Artemis endonuclease complex and the interaction of Artemis with other NHEJ components including DNA Ligase IV.

To fulfil the objectives, various approaches have been taken:

- Bioinformatics analysis of Artemis to identify the possible region of Artemis that interacts with DNA-PKcs
- Cloning, protein expression and purification of different constructs of Artemis, DNA-PKcs and DNA Ligase IV.
- Characterisation of the protein constructs purified including enzyme assays, and biophysical analysis.
- Initiation of fragment-based drug discovery including screening of the fragment library to target the interaction between DNA Ligase IV and Artemis.
- EM studies, including preliminary negative staining screening of DNA-PKcs and Artemis and subsequent cryo-EM studies on different constructs of the DNA-PKcs/Artemis complex
- Analysis based on the cryo-EM results.

During my PhD, some of the objectives were modified in response to published work from other groups. The objective most significantly changed is the screening of stable constructs of Artemis nuclease region. A paper came out in the second year of my PhD showing that the disordered C-terminal tail of Artemis folds back to interact with the nuclease region, explaining why It had proven highly difficult to obtain stable constructs of the nuclease region alone. Therefore, instead of reconstructing the nuclease region, more effort has been spent on the full-length Artemis and the interaction with DNA-PKcs.

In addition to the objectives mentioned, I was also involved in the collaboration between the Blundell group and the Strick group on understanding the temporal organisation of NHEJ

using a single-molecule study. The collaboration work was published on *Nature Structural & Molecular Biology* and will be introduced in section 5.3.

1.6 Overall Organization of The Thesis

The thesis is organised in the following way:

- Chapter 2 describes all the material and methods used in the research described in the thesis
- Chapter 3 describes the initial bioinformatics analysis of Artemis that provides insights into protein structure and interactions and guides the design of different Artemis constructs.
- Chapter 4 describes the purification of protein constructs
- Chapter 5 describes the characterisation of the purified proteins.
- Chapter 6 describes cryo-EM structural work regarding the DNA-PKcs/Artemis complexes and DNA-PKcs.
- Chapter 7 discusses the main outcomes of my thesis and gives a perspective of future work on NHEJ

Chapter 2. Material and methods

2.1 Bioinformatics Analysis

2.1.1 Structural alignment & sequence alignment

Structural homologues are searched using HHpred (Homology detection & structure prediction by HMM-HMM comparison). To find the homologues of specific genes, HomoloGene database (https://www.ncbi.nlm.nih.gov/homologene) and PSI BLAST (Position-Specific Iterated Basic Local Alignment Search Tool) (https://blast.ncbi.nlm.nih.gov/Blast.cgi) are used (Altschul *et al.*, 1997; Soding, Biegert and Lupas, 2005). Structural homologues were aligned using the ModSuite station written by Dr Bernardo Ochoa from Blundell group. ModSuite combines two analysis tools: BATON and FUGUEALI (Shi, Blundell and Mizuguchi, 2001). BATON takes in PDB files of homologous proteins of the protein of interest, producing aligned structure files of the homologues. FUGUEALI aligns the protein of interest to the proteins that have been structurally aligned by BATON. Sequence alignment was performed by the Muscle programme built in the SeaView package (Edgar, 2004; Gouy, Guindon and Gascuel, 2010).

2.1.2 Secondary structure prediction & modelling

For the structured region, Jpred was the main programme used to predict the secondary structures of protein constructs (http://www.compbio.dundee.ac.uk/jpred/) (Drozdetskiy *et al.*, 2015). To model specific protein constructs, Modeller was used following the structural alignment (Šali and Blundell, 1993). I-TASSER was also used to predict the structure of the protein constructs in parallel (Zhang, 2008).

2.1.3 Intrinsically disorder analysis

For the unstructured region, IUPRED was applied to the sequence to analyse the level of intrinsic disorder (http://iupred.elte.hu/) (Dosztanyi *et al.*, 2005). To predict the intrinsically disordered region that plays a role in protein-protein interaction, Anchor was used to search for the intrinsically disordered regions that have the potential to go through concerted folding when interacting with binding partners (http://anchor.elte.hu/) (Dosztányi, Mészáros and

Simon, 2009). In addition, DisEMBL (http://dis.embl.de/) (Linding *et al.*, 2003) and FoldIndex (https://fold.weizmann.ac.il/fldbin/findex) (Prilusky *et al.*, 2005) are used to predict the intrinsic disorder and foldability.

2.2 Molecular Biology

2.2.1 Recombinant constructs and plasmids

The recombinant constructs of NHEJ components built and purified in the thesis are listed in the following table, including Artemis, Ku, DNA Ligase IV, DNA Ligase IV complex and PAXX.

Table 1: Summary of recombinant constructs and plasmids

Construct Name	Protein	Residue range		Mutations
		Begin	End	
Art_Q363_GFP	Artemis	1	363	
Art_S385_GFP	Artemis	1	385	
Art_D394_GFP	Artemis	1	394	
Art_P426_GFP	Artemis	1	426	
Art_Q363_His	Artemis	1	363	
Art_S385_His	Artemis	1	385	
Art_D394_His	Artemis	1	394	
Art_P426_His	Artemis	1	426	
Art_363-399	Artemis	363	399	
Art_385-413	Artemis	385	413	
Art_399-408	Artemis	399	408	
Art_399-426	Artemis	399	426	
Art_413-426	Artemis	413	426	
Art_FL	Artemis	1	692	
Art_FL_Dead	Artemis	1	692	H115A
Ku80_CTD	Ku80	593	732	
Lig4_240	DNA Ligase IV	1	240	
Lig4_230	DNA Ligase IV	1	230	
Lig4_609	DNA Ligase IV	1	609	
LX4_WT	DNA Ligase IV	1	911	
	XRCC4	1	334	
LX4_Dead	DNA Ligase IV	1	911	K273A
	XRCC4	1	334	
PAXX	PAXX	1	204	

Construct Name	Tag (N/C Terminus)	Plasmid	Comments
Art_Q363_GFP	Intein-GFP (C)	p438	codon optimised for insect cell
Art_S385_GFP	Intein-GFP (C)	p438	codon optimised for insect cell
Art_D394_GFP	Intein-GFP (C)	p438	codon optimised for insect cell
Art_P426_GFP	Intein-GFP (C)	p438	codon optimised for insect cell
Art_Q363_His	8His (C)	p438	codon optimised for insect cell
Art_S385_His	8His (C)	p438	codon optimised for insect cell
Art_D394_His	8His (C)	p438	codon optimised for insect cell
Art_P426_His	8His (C)	p438	codon optimised for insect cell
Art_363-399	8His-GST (N)	pGAT3	codon optimised for E.coli
Art_385-413	8His-GST (N)	pGAT3	codon optimised for E.coli
Art_399-408	8His-GST (N)	pGAT3	codon optimised for E.coli
Art_399-426	8His-GST (N)	pGAT3	codon optimised for E.coli
Art_413-426	8His-GST (N)	pGAT3	codon optimised for E.coli
Art_FL	8His (C)	p438	codon optimised for insect cell
Art_FL_Dead	8His (C)	p438	codon optimised for insect cell
Ku80_CTD	8His-GST (N)		Gift from Dr Qian Wu
Lig4_240	8His-GST (N)	pGAT3	Gift from Dr Takashi Ochi
Lig4_230	8His-GST (N)	pGAT3	Modified from Lig4_240
Lig4_609	6His-SUMO	pOPINS	Gift from Dr Takashi Ochi
LX4_WT	6His (C)	pRSFDuet1	Gift from Dr Takashi Ochi
LX4_Dead	6His (C)	pRSFDuet1	Modified from LX4_WT
PAXX	6His (N)	pGAT3	Gift from Dr Takashi Ochi

2.2.2 Oligonucleotides

The following table includes forward and reverse primers for building recombinant constructs and the hairpin DNA (YM164) for nuclease activity analysis purified by desalting and made by Sigma-Aldrich. YM164 is labelled with Cyanine3 on the 5' end.

Table 2: Summary of oligonucleotides

Construct Name	Primer Sequence	Vector
p438_Artemis_F	ATCCAGTGCTCTTCC ATGTCCTCATTTGAAGGGCAGATG	p438
Art_Q363_intein_R	ATCTCCCGTGATGCA CTGGCTTGACCGACACAGTGGTTTG	p438
Art_S385_intein_R	ATCTCCCGTGATGCA GCTGTCTCGGTGCACAGTCCGGGCT	p438
Art D394_intein_R	ATCTCCCGTGATGCA GTCAAACAGATAATCGTCCTCTTCC	p438
Art_P426_intein_R	ATCTCCCGTGATGCA TGGCTGCTTCTCAGACACAGCGGTC	p438
Art_Q363_His_R	TTATCCACTTCCAAT TTAATGGTGATGGTGATGGTGCTGG CTTGACCGACACAGTGGTTTG	p438
Art_S385_His_R	TTATCCACTTCCAAT TTA ATGGTGATGGTGATGGTGGCTG TCTCGGTGCACAGTCCGGGCT	p438
Art_D394_His_R	TTATCCACTTCCAAT TTAATGGTGATGGTGATGGTGGTC AAACAGATAATCGTCCTCTTCC	p438
Art_P426_His_R	TTATCCACTTCCAAT TTA ATGGTGATGGTGATGGTGTGG CTGCTTCTCAGACACAGCGGTC	p438
Art_363-399_F	TTCCAGGGTTCCATGCAGTCTACCGAACCGAAAT	pGAT3
Art_363-399_R	CGAGGTCGACGAATT TTAAATCGGCAGCGGGTCATCGAA	pGAT3
Art_385-413_F	TTCCAGGGTTCCATG AGCGAAGAAGAAGATGACTACC	pGAT3
Art_385-413_R	CGAGGTCGACGAATT TTA CGGGTGAAAGGTTTCCGGATACG	pGAT3
Art_399-408_F	TTCCAGGGTTCCATG ATTCCGCTGCGCCATAAAG	pGAT3
Art_399-408_R	CGAGGTCGACGAATT TTA CGGATACGGAACTTTATGGCGC	pGAT3
Art_399-426_F	TTCCAGGGTTCCATG ATTCCGCTGCGCCATAAAG	pGAT3
Art_399-426_R	CGAGGTCGACGAATT TTA CGGCTGTTTTTCAGAAACTG	pGAT3
Art_413-426_F	TTCCAGGGTTCCATGCCGGAAGTGTTCAGCATGACGGCAGTTT CTGAAAAACAGCCG	pGAT3
Art_413-426_R	CGAGGTCGACGAATTTTACGGCTGTTTTTCAGAAACTGCCGTC ATGCTGAACACTTCCGG	pGAT3
Art_FL_Dead_QC_F	TCCTGCCAGCAGGAGCCTGTCCAGGTAGCGT	p438
Art_FL_Dead_QC_R	ACGCTACCTGGACAGGCTCCTGCTGGCAGGA	p438
LigIV_P231Stop_F	AGGCAACTGCATGATTAATCTGTAGGACTCAGT	
LigIV_P231Stop_R	ACTGAGTCCTACAGATTAATCATGCAGTTGCCT	
YM164	TTTTTGATTACTACGGTAGTAGCTACGTAGCTACTACCGTAGTAAT	

2.2.2 PCR for Ligation Independent Cloning (LIC)

PCR of specific DNA segments was done using Q5 High-Fidelity DNA polymerase (NEB). Generally, a PCR reaction has a total volume of 50ul of: 1x Q5 buffer, 200uM dNTPs mixture,

0.5uM forward primer, 0.5uM reverse primer, 20ng template DNA and 1U Q5 High-Fidelity DNA Polymerase. The protocol of thermocycler used for the amplification is shown in the following table.

Table 3: Protocol of thermocycler for PCR

Initial Denaturation	98°C	30 s
35 cycles	98°C	10 s
	*50-72°C	30 s
	72°C	30 s/kb
Final extension	72°C	2 mins

*The annealing temperature is calculated by NEBTm Calculator (http://tmcalculator.neb.com/ #!/main)

After amplification, the PCR product was applied to 1% agarose gel, cast with TBE buffer and EtBr (Thermo Fisher Scientific), at 100V for 50 minutes to check the outcome of the PCR reaction. The target DNA band would then be extracted from the gel using the QIAquick Extraction Kit (Qiagen) following the protocol. There was one modification to point out that, instead of using 50ul of the elution buffer (Buffer EB) provided in the kit, 30ul of Milli-Q water was used for the final elution of DNA.

2.2.3 Ligation Independent Cloning (LIC)

To insert the amplified target DNA fragments into the linearlised vector, In-Fusion HD Cloning Plus kit (TaKaRa Bio) was used. The total reaction was 10ul including 2ul of In-Fusion enzyme and 8ul of linearlised vector and insert fragment mix. The amount and ratio of the linearlied vector and insert fragment was calculated via the In-Fusion molar ratio calculator (https://www.takarabio.com/learning-centers/cloning/in-fusion-cloning-tools/in-fusion-molar-ratio-calculator). The mixture was incubated at 50°C for 15 minutes for the reaction to complete. The reaction mixture was then stored at -20°C or directly transformed into bacteria cells.

2.2.4 Site-directed mutagenesis PCR

To mutate specific amino acid residues for mutation analysis or introduce a new stop codon for new truncated constructs, site-directed mutagenesis was done using the QuikChange Site-

Directed Mutagenesis Kit (Agilent Technologies). The total reaction volume was 50ul of 1x reaction buffer, 200uM dNTP, 50ng dsDNA template, 125ng forward primer, 125ng reverse primer and 2.5U *PfuTurbo* DNA polymerase. The protocol of thermocycler used for the amplification is shown in the following table.

Table 4: Protocol of thermocycler for site-directed mutagenesis

1 cycle	95°C	30s
18 cycles	95°C	30 s
	55°C	1 min
	68°C	1 min/kb of plasmid

Once all the listed cycles are completed, the reaction mixture is added with 0.5ul of DpnI (20u/ul) and incubated at 37°C for 60 minutes to digest the starting unmutated plasmid templates.

2.2.5 Transformation of bacteria and plasmid amplification

For the amplification of the target plasmid, it was first transferred into bacteria cell. In most cases, Subcloning Efficiency™ DH5α Competent Cells (Invitrogen) were used following the protocol. Specifically, 10ng of DNA plasmid was first added to 50ul of competent cells in a 1.5ml microcentrifuge tube. The mixture was first incubated on ice for 30 minutes and later receives heat shock for 20 seconds in a 42°C water bath. The tube was then placed on ice for 2 minutes, after which 450ul of super optimal broth with catabolite repression (SOC) medium was added to the mixture for another 60-minute incubation in a 37°C shaker. After the recovery, 100ul of the cell culture was plated onto the previously prepared agar plate, which had the specific antibiotics added according to the target plasmid resistance, in a 37°C incubator overnight for selection of the successfully transferred bacteria cells. The plate was then used for colony picking for plasmid amplification and sealed with parafilm and stored under 4°C.

To amplify the plasmid, a single colony collected from the selection plate was added to 5ml of lysogeny broth (LB) for culture in a 37°C shaker overnight. The LB was previously prepared with the specific antibiotics relevant to the target plasmid. After overnight incubation, the cells were collected and the target plasmid was extracted using the QIAprep Spin Miniprep

Kit (Qiagen) following the protocol. There was one modification to point out that, instead of using 50ul of the elution buffer (Buffer EB) provided in the kit, 30ul of Milli-Q water was used for the final elution of DNA.

2.3 Protein Sample Expression and Purification

2.3.1 Protein expression in *E.coli*

Commercial BL21(DE3) Competent *E. coli* cells (New England Biolabs) were transferred with the target plasmid with the protein construct previously built in, following the relevant protocol. A colony from the selection plate was picked and added to the preculture, Terrific Broth (TB) or LB, with the relevant antibiotics. The preculture was then incubated in the 37°C shaker overnight for the later large-scale expression in TB or LB. The volume of the preculture was dependent on the volume of the later large-scale expression. Generally, 10 ml of preculture was required to inoculate 1L of media with the required antibiotics. The inoculated large-scale cultures were then incubated in the 37 °C shaker until the OD600 value reached 0.6. At this point, there were two options for further induction of protein expression. First, isopropyl β-D-1-thiogalactopyranoside (IPTG) could be added to the media to reach a final concentration of 1mM and the cultures remained in the 37 °C shaker for another 3 hours for protein expression. Second, the media could be left in the 16°C shaker for the media temperature to drop back to 16°C. Then IPTG could be added to the media to reach a final concentration of 1mM and the cultures remained in the 16 °C shaker overnight for protein expression.

After the incubation for protein expression, the cultures were centrifuged at 4,200 rpm for 20 minutes at 4°C using the JS-4.2 rotor and J6-MI centrifuge (Beckman Coulter) for cell collection. The supernatant was discarded and the resulting pellets of cells were either directly lysed for further protein purification or flash frozen in liquid nitrogen and stored at -80 °C.

2.3.1 Protein expression in insect cell

Commercial Sf9 cells (Gibco®) were used as the expression system of insect cell. To start with, the recombinant constructs built in the p438 plasmid were transformed into the commercial MAX Efficiency DH10Bac competent E. coli cells (Invitrogen) to produce bacmid following its protocol. The bacmid was then analysed by PCR, using M13 forward and M13 reverse primers, to see if the recombinant construct is inserted into the bacmid. The successfully reconstructed bacmids were then used for generating viruses to transfect insect cells following the protocol of Bac-to-Bac™ Baculovirus Expression System (Gibco®). Generally, to generate baculovirus (P1 virus) for future insect cell transfection, the bacmids obtained previously were transfected into Sf9 cells in Serum-Free Media (SFM) with Cellfectin Reagent. After incubation for 4h at 27°C, the media was replaced with complete growth media and incubated for another 5 days. The baculovirus was harvested. The P1 virus was then amplified with 25ml of sf9 cells at a concentration of 2 million cells/ml to produce P2 virus. The P2 virus was then tested in a growth inhibition assay, where 1ml of 5x, 25x or 125x dilution of P2 virus was added to 25ml of cell culture at a concentration of 1 million cells/ml, to calculate the titer of virus. After calculating the titer, a 3-4-day time-course small-scale expression test was carried out to test the optimal expression time after virus addition to the system. Basically, 25ml of cell culture at a concentration of 2 million cells/ml was transfected by the P2 viruses with a MOI value of 2. Everyday 2 ml of cells were collected. They were later lysed and analyzed by running 5%-12% SDS-PAGE gel. The big-scale insect cell expressions were incubated for protein expression based on the expression test. After the incubation for protein expression, the cultures were centrifuged at 4,000 rpm for 15 minutes at room temperature in a JLA8.100 rotor (Beckman Couter) for cell collection. The supernatant was discarded and the resulting pellets of cells were either directly lysed for further protein purification or flash frozen in liquid nitrogen and stored at -80 °C.

2.3.2 Cell lysis

For the cell lysis, sonicator was mainly used. Cell pellets were thoroughly resuspended in specific cell lysis buffer based on the purification protocol of different protein constructs with cOmplete Mini, EDTA-free protease inhibitor cocktail (Roche). Usually, 30ml- 40ml of the resuspended mixture in the 50ml Falcon tube was lysed each time with the Vibra-Cell VCX130

ultra-sonicator (Sonics & Materials Inc) using the amplitude of 60%- 70% for a total of 1-2 minutes using the pulse cycle of every 10-second sonication followed by 20-second break to limit the heat generated during sonication. After sonication, the lysate was balanced accurately in the Nalgene™ Oak Ridge High-Speed PPCO Centrifuge Tubes (Thermo Fisher Scientific), which were then centrifuged at 30,000 × g at 4 °C for 45 minutes. After the centrifugation, the supernatant was collected and filtered using the 0.2 µm Sartorius™ Minisart™ Single use filter unit (Sartorius™) for the further purification.

2.3.3 Nickel affinity purification

For the nickel-affinity purification, two products were mostly used including Ni-NTA Agarose resins (Qiagen) and HisTrap™ High Performance columns (GE Healthcare). Ni-NTA Agarose resins were mainly used for insect cell purification or small-scale expression test and HisTrap™ High Performance columns were used for the rest. Generally, the nickel affinity purification could be separated into three steps- initial equilibration, sample application and wash and elution. For the initial equilibration of the column, 5-10 column volume (CV) of initial buffer was used to equilibrate the column. The composition of the initial buffer depended on the lysis buffer of the cell of specific protein constructs. The initial buffer should have the same pH and salt concentration as the previous cell lysis buffer with extra imidazole at the final concentration of 15mM. Also, the reducing agent of the initial buffer was either β-mercaptoethanol (BME) at the final concentration of 2mM or tris (2-carboxyethyl) phosphine (TCEP) at the final concentration of 1mM. After the initial equilibrium, the cell-lysis supernatant was loaded to the column, followed by the column washing where the column was washed with at least 5-10 CV of washing buffer which is same as the initial buffer in most of the cases. The final elution step varied based on the concentration change of imidazle. It could be linear-gradient elution, stepwise elution or one-step elution using 500mM imidazole. This should be based on the purification protocol of different protein constructs.

The eluted His-tag protein fraction was usually dialysed at 4°C overnight to reduce the amount of imidazole. This step could be combined with TEV protease to cleave the His tag. The ratio between the amount of TEV protease added and the amount of target His-tag protein was 1:100. After dialysis and cleavage, the sample was filtered using the 0.2µm Sartorius™

Minisart™ Single use filter unit (Sartorius™). The filtered sample was then loaded to the preequilibrated HisTrap™ High Performance column for a "reverse" His-tag purification to get rid of the uncut His-tag protein. The flowthrough of the column was collected for the following purification steps.

2.3.4 Glutathione S-transferase (GST) affinity purification

GSTrap HP columns (GE Healthcare) were mainly used for the GST affinity purification. The whole procedure is similar to Ni affinity purification mentioned above. The GSTrap HP column was first equilibrated with the 5-10 CV initial buffer, which is similar to the cell-lysis buffer without the cOmplete Mini, EDTA-free protease inhibitor cocktail (Roche). The cell-lysis supernatant was then loaded to the column followed by column washing with 5-10 CV initial buffer. Unlike Ni affinity purification, the final elution of GST affinity purification was usually one-step elution due to the high specificity of the GST tag. During this step, the column was loaded with elution buffer which is the same as the initial buffer with extra L-reduced glutathione at the concentration of 25mM. The eluted fraction was usually treated in the similar way to that of the Ni affinity purification with dialysis, TEV cleavage, filtering and "reverse" GST-tag purification. The flowthrough was then collected for the coming purification steps.

2.3.5 Ion Exchange Chromatography

For the ion exchange chromatography, various columns were used including HiTrap Q HP (GE Healthcare), HiTrap SP HP (GE Healthcare), Mono Q (GE Healthcare), Mono S (GE Healthcare), HiTrap Heparin HP (GE Healthcare) and hydroxyapatite column (Biorad). Generally, ion exchange chromatography was also separated into three steps- equilibration, sample application and wash and elution. To start with, the column was equilibrated with the initial buffer of low salt, meaning low ionic strength. The sample to be loaded was also previously dialysed in the same buffer of low salt. The sample at low salt was then applied to the equilibrated column for binding. The column was then washed with 5-10 CV of the equilibrium buffer. The elution could be stepwise elution and linear elution. In most of the cases, linear elution was used. There were two buffers- the buffer with low salt concentration (eg. 10mM NaCl) and the buffer with high salt concentration (eg. 1M NaCl). The protein was eluted on an

ÄKTA system (GE Healthcare) using the linear gradient from 100% low-salt buffer to 100% high-salt buffer within the elution volume of 20 CV. To point out, the hydroxyapatite column (Biorad) is incompatible with Tris and EDTA, which should be avoided in the buffer.

2.3.6 Size Exclusion Chromatography

For the size exclusion chromatography, considering the size of the protein construct and the amount of protein sample, different columns were used including Superdex/ Superose columns. Previously collected elution fraction containing the target protein was first concentrated using proper Vivaspin spin concentrators (Sartorius) to the smaller volume, which is usually the 1% CV of the gel filtration column. The sample was then centrifuged at 13000x g for 15 minutes at 4°C to accelerate the precipitation of aggregate and insoluble contamination in the sample. Later, the centrifuged sample was added to the specific gel filtration column, which was preequilibrated with the buffer designed as the final storage buffer for the purified protein. The gel filtration was performed using the ÄKTA system (GE Healthcare).

2.4 Protein Biochemical and Biophysical Analysis

2.4.1 Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

To analyse the size of the protein and visualise the protein sample, protein samples were applied to SDS-PAGE. Before added to the gel, protein samples were first mixed up with 4X loading buffer (200 mM TRIS pH 6.8, 10% SDS, 0.05% bromophenol blue, 20% glycerol, and 700 mM β-mercapto-ethanol) and incubated on the 98°C Mini Dry Bath Incubator (Starlab) for 10 minutes for denaturation.

Two types of SDS-PAGE gels were used for analysis- commercial SDS-PAGE gels including the NuPAGE™ 4-12% Bis-Tris Protein Gels (Invitrogen) and the homemade SDS-PAGE gels. In the case of using the NuPAGE™ 4-12% Bis-Tris Protein Gel, samples were added and run at 200V for 50 minutes in NuPAGE MOPS SDS running buffer (Invitrogen) for medium- to large-size proteins or run at 200V for 30 minutes NuPAGE MES SDS running buffer (Invitrogen) for small-

size proteins. As for the homemade SDS-PAGE gels, they were usually run at 150V for 70 minutes. The recipe of the homemade SDS-PAGE gel is as follow:

Table 5: Recipe of homemade SDS-PAGE gel

Reagent	per 15% running gel	per 5% stacking gel
Milli-Q water	1.1ml	950µl
Acylamide (30%)	2.5ml	225ul
1.5M Tris-HCl pH8.8	1.25ml	\
0.5M Tris-HCl pH6.8	\	400µl
10% SDS	100µl	17µl
25% Ammonium persulfate	40ul	7µl
TEMED	10µl	2ul

Once the electrophoresis was completed, the gel could be stained with Coomassie Blue staining (Expedeon) or silver staining using the Pierce™ Silver Stain Kit (Thermo Fisher Scientific) following the protocol.

2.4.2 Nickel Affinity Pull-Down Assay

The nickel affinity pull-down assay was used to identify the fragments of Artemis C-terminal peptides interacting with DNA-PKcs. The resin used in the assay was Ni-NTA Agarose (Qiagen) and the initial/wash buffer was: 20mM HEPES pH7.5, 200mM NaCl, 15mM imidazole and 2mM MBE. 5ul of the beads were used in one pull-down experiment group, equilibrated with the initial/wash buffer. After equilibration, 1ul of Artemis peptide at the concentration of 75uM and 4ul of DNA-PKcs at the concentration of 1uM were added to the beads for incubation at 4°C for 60 minutes. The system was then washed three times with the initial/wash buffer and eluted using the elution buffer which was similar to the initial/wash buffer but with 300mM imidazole instead. The elution fraction was added with the 4X loading buffer, denatured and applied to the NuPAGE™ 4-12% Bis-Tris Protein Gel (Invitrogen) for running at 200V for 30 minutes under 1x NuPAGE MES SDS running buffer (Invitrogen). The gel was last stained using the Pierce™ Silver Stain Kit (Thermo Fisher Scientific).

2.4.3 Ligation Assay

Ligation assay was used to test the purified DNA ligase IV complex. To do the assay, 50 ng of the linearlised plasmid was mixed with the DNA ligase IV complex, with the final concentration of 100nM, in a 20µl reaction volume containing 25mM Tris-HCl (pH 7.5), 150 mM KCl, 1 mM DTT, 10µg/ml BSA. The mixture was then incubated at 37°C for 30 minutes. To stop the reaction, the mixtures were incubated at 50°C for another 30 minutes after adding 2µl of the stop buffer (100 mM EDTA, 0.1% (w/v) SDS) and 0.2µl of 20 mg/ml Proteinase K. Reaction mixtures were applied onto 0.8% agarose gel in 1xTBE buffer for electrophoresis (run in room temperature at 70 V for 90 minutes). The gel was stained with SYBR Gold and visualized using a UV imager

2.4.4 Nuclease Assay

Nuclease assay was done to test the activity of Artemis endonuclease activity on DNA hairpin. To do the assay, a mixture of hairpin DNA labelled with cyanine dye 3 at the final concentration of 20 nM, Artemis at the final concentration of 50 nM, and DNA-PKcs at the final concentration of 50 nM and other specified proteins at certain concentrations were incubated in 20µl of nuclease buffer composed of 5 mM Tris-HCl pH 7.5, 10 mM KCl, 0.5 mM ATP and 1 mM DTT. After the incubation at 37°C for 60 minutes, 20µl of stop buffer containing 95% formamide, 18 mM EDTA, 0.25% SDS was added and the mixture was treated with heat shock at 98°C for 5 minutes. The samples were then quickly cooled on ice and separated in a 12% denaturing polyacrylamide gel in TBE buffer by electrophoresis at 150 V for 15 minutes. After running, the gel was visualised using Typhoon™ FLA 9000 (GE healthcare).

2.4.5 Circular Dichroism (CD)

Circular dichroism was used to test whether the purified full-length Artemis sample was folded. The experiment was conducted using the Aviv Model 410 (Aviv Biomedical) at the Biophysics facility of the Department of Biochemistry, University of Cambridge under the help of Dr Katherine Stott. Sample preparation and the protocol was referred to the SOP provided on the facility website (http://www.biophysics.bioc.cam.ac.uk/?page_id=37).

2.4.6 Dynamic Light Scattering (DLS)

Dynamic light scattering was used to test the molecular size and sample homogeneity for cryo-EM preparation. The experiment was conducted using the Malvern Zetasizer Nano S (Malvern Panalytical) at the Biophysics facility of the Department of Biochemistry, University of Cambridge under the help of Dr Katherine Stott. Sample preparation and the protocol was referred to the SOP provided on the facility website (http://www.biophysics.bioc.cam.ac.uk/?page_id=37).

2.4.7 Mass Spectrometry (MS)

Mass spectrometry was mainly used to check if the purified protein was the target construct. Protein samples were first run on SDS-PAGE gel and stained using Coomassie Blue staining (Expedeon). The stained SDS-PAGE gel was washed with Milli-Q water and stored in 5% methanol solution and submitted to the CCPcore mass spectrometry facility (University of Cambridge, Department of Biochemistry) for subsequent band cutting, trypsin digestion and MALDI-TOF mass fingerprinting.

2.4.8 Differential Scanning Fluorimetry (DSF)

Differential scanning fluorimetry or thermal shift assay (TSA) was used to test the binding of fragments with DNA Ligase IV. The experiment was conducted using the Bio-Rad CFX machine (BioRad) at the Biophysics facility of the Department of Biochemistry, University of Cambridge. It was carried out in a 96-well plate. Each well contains 25µl of reaction mixture of DNA ligase IV at the final concentration of 10µM, saturated fragments in DMSO and 5x Sypro orange dye. Appropriate positive (Protein, DMSO and AdoMet) and negative (Protein, DMSO only) controls were also included. The measurement was performed using certain thermal cycler with 25°C for 10 minutes followed by a linear increase of 0.5°C every 30 seconds until reaching the final temperature of 95°C. The results were analysed using Microsoft excel sheet.

2.4.9 Protein crystallisation

Both commercial and manual crystallisation screens were applied to search for the crystallisation condition for DNA ligase IV constructs. The commercial screens were set up with the 96-well 2-drop MRC plates (Swissci) using the Mosquito crystallisation robot (TTP

Labtech) in the sitting drop manner. One well contains 0.2µl of protein solution and 0.2µl of the crystallisation condition buffer. The completed plates were then stored at 19 °C in a Rock Imager (Formulatrix) for routine light imaging and UV absorbance imaging. The manual screens were set up with the 24-well VXD plates (Hampton Research) in the hanging drop manner. One well contains 500µl of the crystallisation condition buffer and one hanging drop which contained 1µl of protein solution and 1µl of the crystallisation condition buffer. The plates were stored at 16°C in a designated crystallisation room and inspected under microscope manually on a daily basis.

2.5 Electron Microscopy Sample Preparation, Data Collection and Analysis

2.5.1 Grid preparation (Negative stain & cryo-EM)

For negative staining, CF400-CU grid (Electron microscopy sciences) was glow discharged using PELCO easiGlow GlowDischarge system (Ted Pella Inc., USA) for 30 seconds with the 15mA current. After glow discharge, 2µl of protein sample was applied to the and incubated for 30 seconds. The sample was then blot off using filter paper. The grid surface was later washed via touching the surface of a Milli-Q water drop and then blot off with filter paper. The operation was repeated with another fresh drop of Milli-Q water. The grid surface was then put in contact with a drop of 1% uranyl acetate solution for 20 seconds, blot off with filter paper and left for airdry.

For cryo-EM, different grids were used, including QUANTIFOIL R 1.2/1.3 300 mesh Cu/Au (Quantifoil); UltraAuFoil 1.2/1.3 300 mesh (Quantifoil). Grids were glow discharged using PELCO easiGlow GlowDischarge system (Ted Pella Inc., USA) for 60 seconds with the 25mA current. The glow discharged grids were then frozen in liquid ethane using FEI Vitrobot Mark IV (Thermo Fisher Scientific) under 4°C and 100% relative humidity. The blot force and blot time varied among different samples, buffer conditions and grid types. Generally, the blot force was between -5 and +5 and the blot time was between 2 and 4 seconds.

2.5.2 Sample screening and data collection

Negative staining was merely used for screening of the sample to check if the sample was homogeneous. FEI Tecnai F20 (Thermo Fisher Scientific) of the Wolfson Electron Microscopy Suite of the Department of Materials Science and Metallurgy, University of Cambridge was used for negative staining screening under the help of EM specialists.

Cryo-EM was used for both sample screening and data collection. Different microscopes were used for cryo-EM experiments including 300 kv Titan Krios I at eBIC Diamond Light Source (Thermo Fisher Scientific), 200 kv Talos Arctica at Department of Biochemistry, University of Cambridge (Thermo Fisher Scientific) and 300kv Titan Krios at Department of Biochemistry, University of Cambridge (Thermo Fisher Scientific). Different parameters were used for different data collections and the details will be listed in the result chapters.

2.5.3 Data analysis

Collected cryo-EM datasets were analysed mostly by Relion 2.0/3.0. Other packages were also tried including cryo-SPARC, Scipion and WARP. After the 3D reconstructions and refinements, cryo-EM maps were visualised and analysed with Chimera. Some of the maps were also modelled using modelling software including Phenix and CCPEM. Detailed process of data analysis will be described in the result chapters.

Chapter 3. Bioinformatics analysis of Artemis

A bioinformatics analysis of Artemis was first carried out to guide the later experimental biophysical and biochemical studies. Previous research had proposed that the nuclease region of Artemis extended from residue 1 to 385 but the remaining sequence comprising the C-terminal region of the protein, referred to here as the C-terminal tail, was suggested from functional and biochemical dissection to be flexible without ordered structure (Pannicke *et al.*, 2004). For a more detailed analysis, including more certain identification on the nuclease region, a series of sequence alignments including structural alignment, secondary structure prediction and modelling were carried out. Moreover, intrinsic disorder analysis was used to investigate the disordered region of Artemis, with the objective of providing insights of further possible protein-protein interaction sites similar to the Artemis-DNA Ligase IV interaction site.

3.1 Sequence Alignments for the Nuclease Domain

Different sequence alignments were used to analyse the possible boundary of Artemis nuclease region taking into account the available information of structures of the nuclease region of the homologues, secondary structure prediction and the conservation of amino acid residues.

Initially, following the HHpred analysis, SNM1A and SNM1B, which had structures available, were chosen as homologues for structural alignment and comparison, as they aligned with the highest probability and lowest e value. SNM1A is a 5'-3' exonuclease and might be involved in DNA inter-strand crosslink repair (Sengerová *et al.*, 2012). SNM1B, also named as Apollo, the twin brother of Artemis in the ancient Greek religion and myth, is a 5'-3' exonuclease playing a key role in telomere protection and maintenance during S phase (Liu *et al.*, 2009; Sengerová *et al.*, 2012). The PDB files of SNM1A and SNM1B used for the alignment (5AHR and 5AHO) contained only the nuclease domains of those proteins (700-1040 for SNM1A and 1-335 for SNM1B) (Allerston *et al.*, 2015). Although the three proteins are homologues, the sequence identities among SNM1A 700-1040, SNM1B 1-335 and Artemis

1-385 are not high, ranging from 26% to 32% (Table 6). Nevertheless, both structures were applied to the BATON and FUGUEALI analysis for structural alignment.

Protein / Sequence Identity / Protein	SNM1A	SNM1B	Artemis
SNM1A	100%	28%	32%
SNM1B	28%	100%	26%
Artemis	32%	26%	100%

Table 6. Sequence identities of the nuclease regions of the homologues, SNM1A, SNM1B and Artemis. The alignment comparison shows they are distantly related with the sequence identities at the level of ~30%.

Assuming that SNM1A(PDB:5AHR), SNM1B (PDB:5AHO) and Artemis share similar structures as structural homologs, based on the structural alignment of (Figure 16A), the structured nuclease region of Artemis should end around Q363, which is different from the previously proposed S385. Moreover, another series of secondary structure predictions were carried out within the package of Jpred around the transition zone between the nuclease region and C-terminal tail (Figure 16B). Predictions of different methods showed different results. In summary, the region before R360 is predicted to be structured with high confidence consistently, while the region after R360 is predicted either unstructured or structured with very low confidence. In addition, sequence alignment of Artemis homologs from different species was conducted to understand the degree of conservation of residues at the border of the nuclease region and C-terminal tail (Figure 16C). There are some conserved hydrophobic patches within the extended C-terminal region after Q363. Also, there are a few conserved positively or negatively charged patches included in the same region, which may also be relevant to Artemis and DNA-PKcs interaction and will be described in detail. Therefore, considering all the analyzed alignments, four residues, Q363, S385, D394, P426, were predicted to be possible end points of the Artemis nuclease region.

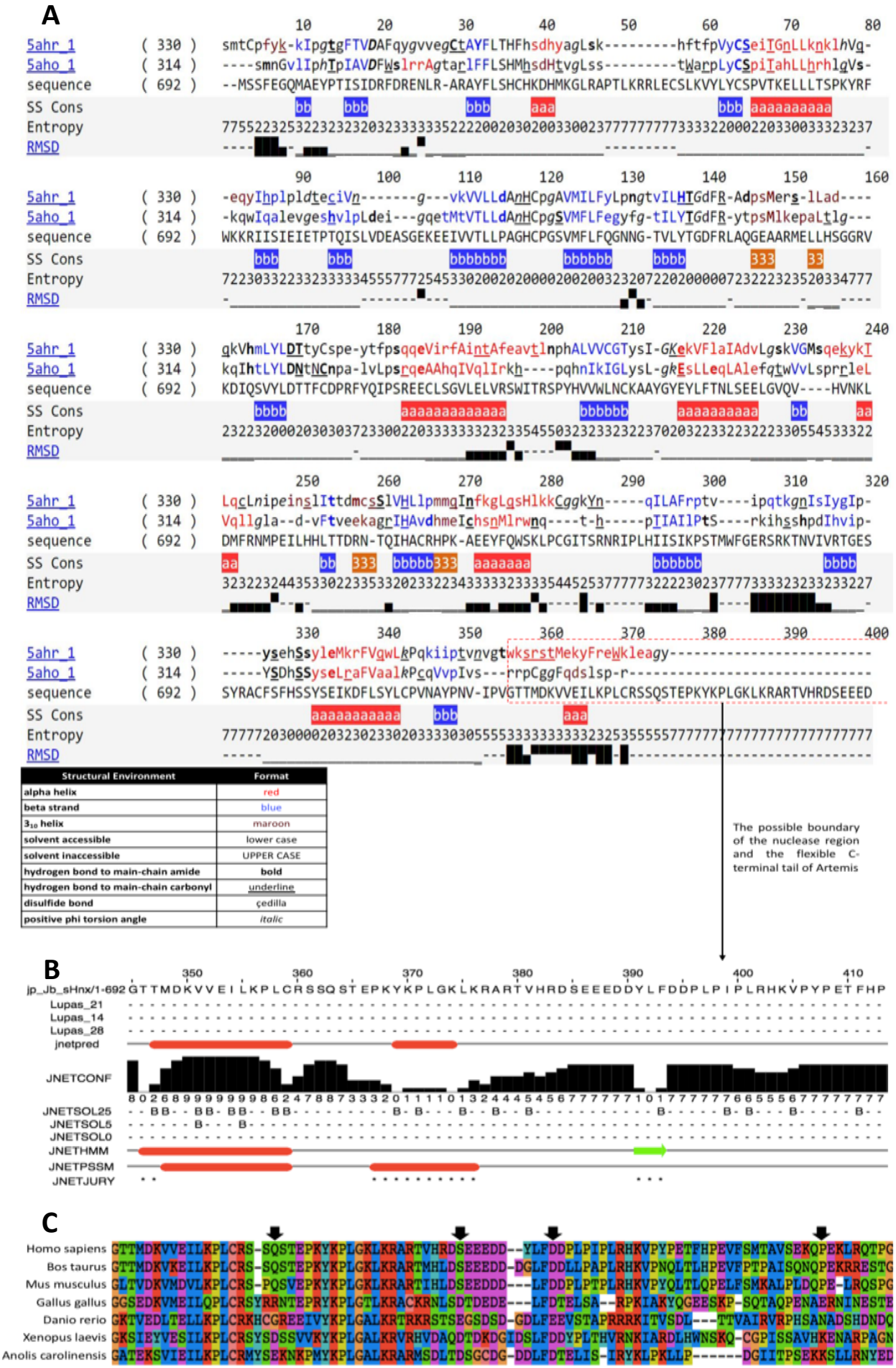

A
B
C
Structural Environment
Format
alpha helix
red
beta strand
blue
3_{10} helix
maroon
solvent accessible
lower case
solvent inaccessible
UPPER CASE
hydrogen bond to main-chain amide
bold
hydrogen bond to main-chain carbonyl
underline
disulfide bond
çedilla
positive phi torsion angle
italic
The possible boundary of the nuclease region and the flexible C-terminal tail of Artemis
Homo sapiens
Bos taurus
Mus musculus
Gallus gallus
Danio rerio
Xenopus laevis
Anolis carolinensis

Figure 16. Bioinformatics analysis on Artemis nuclease region. (A) Structural alignment of SNM1A (PDB:5AHR), SNM1B (PDB:5AHO) and Artemis in JOY format defined below; (B) Secondary structure prediction around the possible boundary of Artemis nuclease region and C-terminal tail using Jpred. Details on acronyms used: Lupas_21/ Lupas_14/ Lupas_28 are coiled-coil predictions for the sequence are binary predictions for each location; JNetPRED is the consensus prediction where helices are marked as red tubes and sheets are marked as dark green arrow; JNetCONF shows the confidence estimate for the prediction; JNetHMM is the HMM profile-based prediction where helices are marked as red tubes and sheets are marked as dark green arrows; JNETSOL25/ JNETSOL5/ JNETSOL0 are predictions of the solvent accessibility at the solvent accessibility cut-offs of 25%, 5% and 0% where 'B' indicates 'buried' and '-' indicates 'exposed'; JNETPSSM is the PSSM-based prediction where helices are marked as red tubes and sheets are marked as dark green arrows; JNETJURY-- '*' means that the JNETJURY was used to rationalize highly inconsistent primary predictions; (C) Sequence alignment of Artemis homologs from different organisms around the possible boundary of Artemis nuclease region and C-terminal tail. The possible positions of the cutoff of nuclease region are labelled with black arrow including Q363, S385, D394 and P426.

3.2 Artemis Nuclease Domain Models and Comparison

In addition to the previous sequence alignment and comparison, modelling of the Artemis nuclease region (Artemis 1-363) was carried out using two packages: Modeller and iTasser (Šali and Blundell, 1993; Zhang, 2008). This was important as it provided a fast and straightforward prediction of the structure of interest, which could be used to compare with the structures of the homologs (SNM1A and SNM1B).

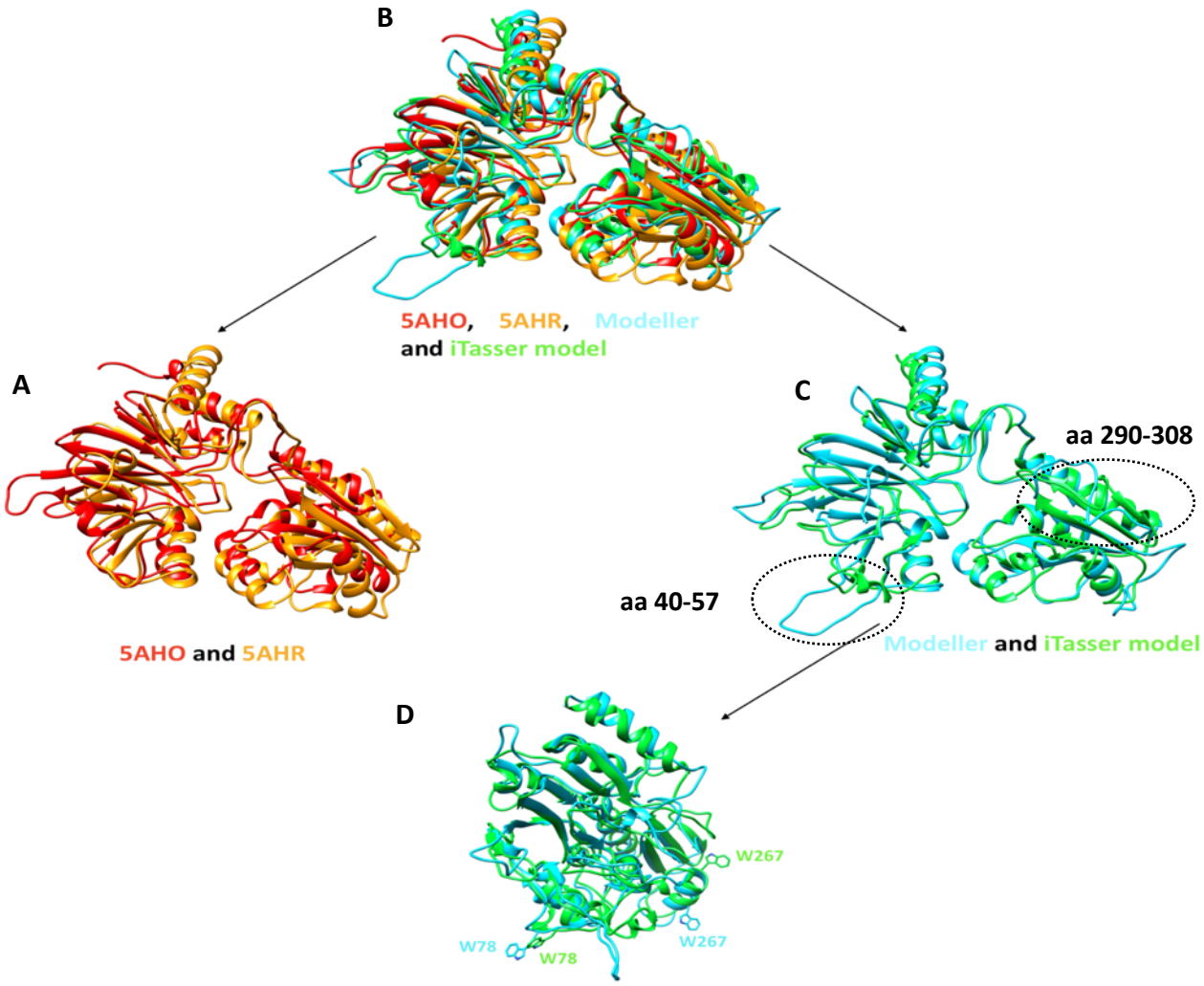


Figure 17. Comparison of the structures of homologs and models of the Artemis nuclease region (Artemis 1-363) using two independent packages of Modeller and iTasser. (A) Structural alignment of SNM1A (pdb: 5AHR; orange colour) and SNM1B (pdb: 5AHO; red colour); (B) Structural alignment of SNM1A (pdb: 5AHR; orange colour), SNM1B (pdb: 5AHO; red colour), Modeller model of Artemis 1-363 (cyan colour) and iTasser model of Artemis 1-363 (green colour); (C) Structural alignment of Modeller model of Artemis 1-363 (cyan colour) and iTasser model of Artemis 1-363 (green colour); (D) Differences between the two models. The aligned models of panel C are rotated 90° and the tryptophan residues (W78 and W267) that differ in position are labelled.

As shown in Figure 17A, the structures of the nuclease region of SNM1A (pdb: 5AHR) and SNM1B (pdb: 5AHO) are very similar, even though the sequence identity is low (Table 6). Compared to the structures of 5AHO and 5AHR, both models of Artemis 1-363 have a similar scaffold but are more flexible as the numbers and lengths of loops are higher (Figure 17B). Due to the flexibility difference, 5AHO and 5AHR are more compact relative to the Artemis models. The two models predicted by different packages are similar but have clear differences. The main differences are within the loops of residues 40-57 and residues 290-308, showing that the flexibility of the molecules is challenging the modelling. In addition to the inconsistent modelling of some loops, there are other differences (Figure 17C). For example, in both models W78 and W267 are placed as exposed residues facing the solvent without any stabilisation from nearby hydrophobic amino acid residues. This is unlikely to be the case and does not happen in 5AHR and 5AHO. There are many possibilities of this situation. For example, there may be other binding partners of Artemis to interact with those exposed hydrophobic residues. However, there may be errors in the alignment or more likely building regions that are absent from the templates used in the comparative modelling. Whatever the reason may be, it is clear that this Artemis nuclease region differs in conformation from those of SNM1A and SNM1B.

3.3 Intrinsic Disorder Analyses of Artemis

As mentioned above, the C-terminal tail of Artemis, comprising almost a half of the protein, has extensive intrinsic disorder. In addition, the previous analysis and modelling of the nuclease region also indicated that the largely structured N-terminal region of Artemis is likely also to have disordered loops. Therefore, it would be interesting to look at the overall intrinsic disorder of the protein. Moreover, the disordered Artemis 485-495 region was shown to go through concerted folding to form a helix when in contact with DNA Ligase IV. Analysis of the disorder in C-terminal tail may be informative for further prediction and experiment to find the similar interaction regions on the Artemis flexible tail.

To investigate the intrinsic disorder of Artemis, three independent packages were used: DisEMBL, FoldIndex and IUPred/Anchor (Figure 18) (Linding *et al.*, 2003; Dosztanyi *et al.*, 2005; Prilusky *et al.*, 2005). Generally, all packages predicted that Artemis is highly disordered with a clear trend that the C-terminal tail has a higher degree of disorder. In the DisEMBL analysis (Figure 18A), three different criteria are taken into consideration including $loops/coils$, $hot\ loops$ and $REMARK-465$.

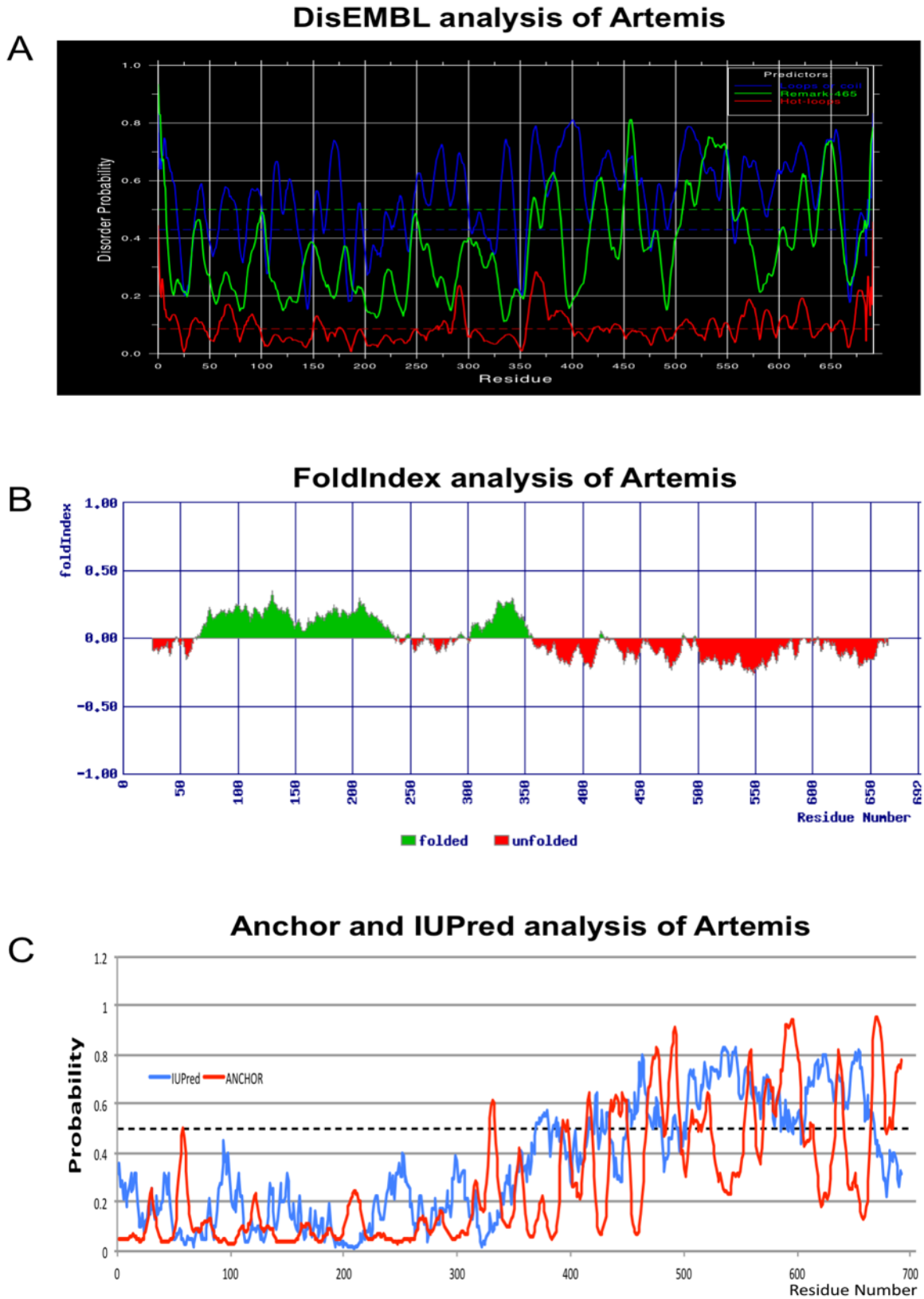


Figure 18. Intrinsic disorder analysis of Artemis. (A) DisEMBL analysis of Artemis including the analysis of loops or coil (blue), Remark 465 (green) and Hot loops (red). The probabilities predicted are shown as curves and the scores should always be compared to the relative random expectation values, which are the dotted lines; (B) FoldIndex analysis of Artemis where the folded region (green) has a positive foldindex and the unfolded region (red) has a negative foldindex; (C) Anchor and IUPred analysis of Artemis. The probabilities are shown as curves and the scores should always be compared to the probability of 0.5, which is the dotted line.

In all three packages, when the disorder probability of a region is higher than the relative random expectation value, which is the dotted line, the region is predicted to be more disordered.

Loops/coils predicts regions that are not necessarily disordered all the time. However, all the disordered regions are found only on loops or coils. Therefore, this could be treated as a necessary but not sufficient requirement for a disordered region. The regions that are predicted not to be loops or coils are located mainly before residue 360, although almost half of the regions before 360 are predicted to be loops or coils. As for the C-terminal tail, only three small regions (around 480, around 560 and around 675) are predicted not to be loops or coils. The Remark 465 analysis shows the region of missing coordinates in X-Ray structures as defined by REMARK-465 entries in PDB. Missing coordinates usually indicate the flexibility of intrinsic disorder and could be used as a way to predict disorder. According to the Remark 465 analysis, the first few amino acids of Artemis are likely to be disordered. Also, most of the C-terminal tail after residue 360 is likely to be flexible except for four regions including regions around 380-420, 440-450, 470-505, 565-625, 655-690 that may fold in the presence of binding partners including other parts of Artemis. As for the Hot loops analysis, it is a subset of the Loops/coils analysis and picks out the regions of loops/coils with high degree of flexibility. Many of the predicted loops or coils regions within the nuclease region are predicted not to be flexible including residue 100-150. However, most of the region before residue 100 is predicted to be flexible. Also, more than half of the C-terminal tail, mostly after residue 500, is predicted to be flexible loops or coils.

Unlike focusing on the intrinsic disorder of the protein, FoldIndex analysis emphasizes whether the protein would fold using the sequence provided. The level of disorder can be estimated as the opposite property of foldability of the region. Positive FoldIndex of the residue indicates the potential of it to fold while negative FoldIndex means unfolded. According to the results (Figure 18B), most of the region before 360 should be folded except for the region around 1-60 and 240-290. As for the C-terminal tail, most was predicted to be unfolded except for two small regions around 413-426 and 490.

The IUPred and Anchor analysis focuses on the prediction of the level of disorder and disordered but foldable binding region (Figure 18C). When the probability is higher than 0.5, it indicates higher level of disorder (IUPred) or higher potential for the disorder region to bind a partner (Anchor). The IUPred showed that the whole region before residue 360 is likely not to be disordered. Furthermore, most of the C-terminal tail should be flexible except that the probabilities of some regions of the C-terminal tail fluctuate around the level of 0.5 including the regions around 390-410, 440-450, 480-490 and 680-692. The Anchor analysis showed that, among the flexible regions with IUPred probability bigger than 0.5, some have the potential to bind a partner. Those include regions around 415-420, 470-480, 490-495, 510-530, 550-560, 565-605 and 630-640.

Although the results of all three analyses are not consistent in all aspects, there are some common points among them. First of all, the nuclease region is more structured compared to the C-terminal tail. However, there are many loops or coils or unfolded regions within the region. They are not necessarily intrinsically disordered, although the nuclease region on its own may be highly flexible. Besides, the C-terminal tail is largely intrinsically disordered while some of the regions may be able to fold or have binding partners. The regions with the mentioned potential are mainly sitting on two sites around residue 420 and 490.

Interestingly, the region around 490 is exactly the site where Artemis interacts with DNA Ligase IV. As for the region around 420, there was no unknown binding partners or structured domain. However, it is not far from the previously proposed site of Artemis (Artemis 399-404) interacting with DNA-PKcs. As there is no identification of the region of Artemis interacting with DNA-PKcs, it could be that the region around 420 is involved in the interaction of Artemis and DNA-PKcs.

To further determine the region of Artemis interacting with DNA-PKcs, in addition to the intrinsic disorder analysis, the sequence analysis around the region 399-420 is also important. According to the sequence alignment (Figure 16C), within the region of 399-404, although the residues L401 and R402 were proposed to be the key residues, only R402 and K404 are conserved in all species. In the flanking region around 399-404, there are two regions that stand out from the others.

The first is the small hydrophobic patch of residues L392 and F393, which are conserved among all species, and the flanking negative-charge patch around E386. There is a trend to increasing negative charge in more evolutionary advanced organisms. It is unclear if the highly conserved hydrophobic patch could be a part of the nuclease region or could be involved in the protein-protein interaction. The second region is between T410 and S422, where the sequences are not highly conserved but reveal a trend of increasing hydrophobicity.

Therefore, to identify the region of Artemis interacting with DNA-PKcs, considering the analysis of sequence alignment and intrinsic disorder, five fragments of the C-terminal region of Artemis with an N-terminal His tag and a following GST tag were purified, including Artemis 363-399; Artemis 399-408; Artemis 385-413, Artemis 399-426 and Artemis 413-426.

3.4 Summary

The nuclease regions of SNM1A and SNM1B were used for the homology modelling of the nuclease region of Artemis. Moreover, with the sequence alignment, the boundary of the nuclease region should sit between Q363 and P426. iTasser was also used for the modelling. The models indicated that the nuclease region of Artemis contains more loop regions compared to SNM1A and SNM1B nuclease region. To point out, the models of Artemis nuclease region have some deficiencies. There may be some binding partners of Artemis or there may be errors in the alignment or more likely building regions that are absent from the templates used in the comparative modelling. In fact, at the beginning first and a half year of my PhD, I tried to build truncated constructs of Artemis based on the bioinformatics analysis for crystallisation to study the structure of the nuclease domain. Strikingly, all the truncated constructs were not stable while the full-length protein, although with a low yield, was purified stably. This led me to the idea of whether the proposed N-terminal nuclease domain interacts with the C-terminal flexible tail for stabilisation. Later, a paper came out during my second year of PhD showed that the C-terminal domain interacts with the N-terminal nuclease domain (Niewolik *et al.*, 2017). It was proposed that the region around residues 456-458 is important for this self-interaction, targeting the residues 1-7 (Niewolik *et al.*, 2017). This may also be an autoinhibition of the nuclease activity and explain why Artemis cannot cut hairpin DNA on its own (Niewolik *et al.*, 2017). Therefore, it may be better to study Artemis at full length rather than truncated recombinants to understand how the nuclease is activated or regulated.

In addition, a series of bioinformatics analysis were conducted targeting the intrinsic disorder of the protein. They were also used to estimate the region of Artemis interacting with DNA-PKcs. According to the results, the region between Q363 and P426, which is also the region where the nuclease region is likely to end, should contain the region interacting with DNA-PKcs. At the end, five fragments of the C-terminal region of Artemis were designed to test the binding with DNA-PKcs, including Artemis 363-399; Artemis 399-408; Artemis 385-413, Artemis 399-426 and Artemis 413-426

Chapter 4. Protein Purification

Protein purification is the basic part of the project. Without purified protein, none of the later biochemical, biophysical and structural studies can be done. During my PhD, protein purification of various constructs has been one of the most challenging and difficult parts of the whole project for many reasons:

First, there were no canonical purification methods for some constructs and optimisation was needed based on different published protocols (Artemis, DNA Ligase IV constructs);

Second, many recombinant constructs are difficult to purify with more than four steps of chromatography and DNA-PKcs is purified from HeLa cells, different from the recombinant constructs (Artemis, DNA Ligase IV constructs);

Third, some proteins have low yield (Artemis, DNA-PKcs, DNA Ligase IV constructs);

Fourth, Artemis full-length protein is flexible and difficult to obtain 100% purity;

Fifth, the HeLa cells for the purification of DNA-PKcs are expensive.

Various constructs of Artemis, DNA-PKcs, DNA ligase IV and Ku were purified. In this chapter, the purification of the proteins is introduced.

4.1 Purification of Artemis Constructs

The purification of Artemis constructs can be separated into two steps, one targeting the full-length constructs and the other targeting the C-terminal peptides within the intrinsically disordered region.

4.1.1 Purification of the full-length Artemis constructs

Two constructs of full-length Artemis were prepared: wild-type Artemis and Artemis H115A. While wild-type Artemis is physiologic and functional, Artemis H115A is catalytically inactive as H115 together with H33 and H35 is involved in coordinating one Zn ion. Previous work showed that the Artemis H115A mutant has little endonuclease activity (Li *et al.*, 2014). The inactive mutant was purified to stabilise and study the conformation when the hairpin DNA is locked at the Artemis catalytic pocket. The wild-type Artemis was used to examine the enzyme properties of the DNA-PKcs/Artemis endonuclease complex.

The His tag is placed at the C terminus of the full-length constructs for two reasons. First, a His tag at the N-terminus could disturb the protein catalytic pocket as the N terminus is close to the active site. Furthermore, as the C-terminal tail is highly unstructured and very likely to be cleaved by proteases, a His tag at the C terminus allows proteins with the C-terminal tail cleaved to be filtered out in the first step of affinity chromatography.

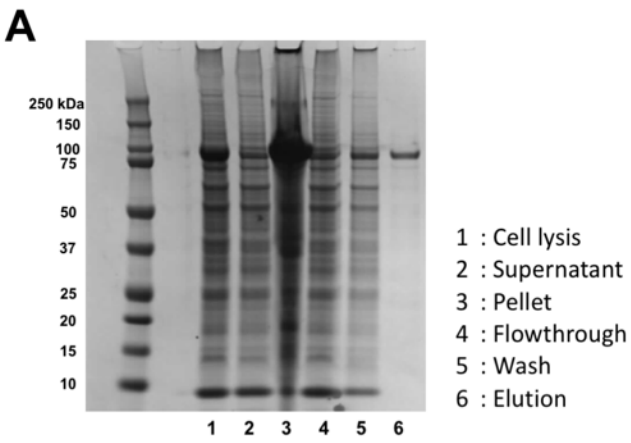
Artemis is so flexible and degradable that the purification has been highly challenging and there is no one canonical purification protocol. My purification of the full-length Artemis constructs was optimised many times based on published procedures. Although it was unlikely to get pure and undegraded Artemis, my method has managed to obtain the functional Artemis and remove most of the heterogeneity according to gel-filtration profile and the results of electrophoresis. In general, the purification can be considered as three steps: His-tag purification, Q-column purification and gel-filtration purification.

4.1.1.1 Purification of wild-type Artemis

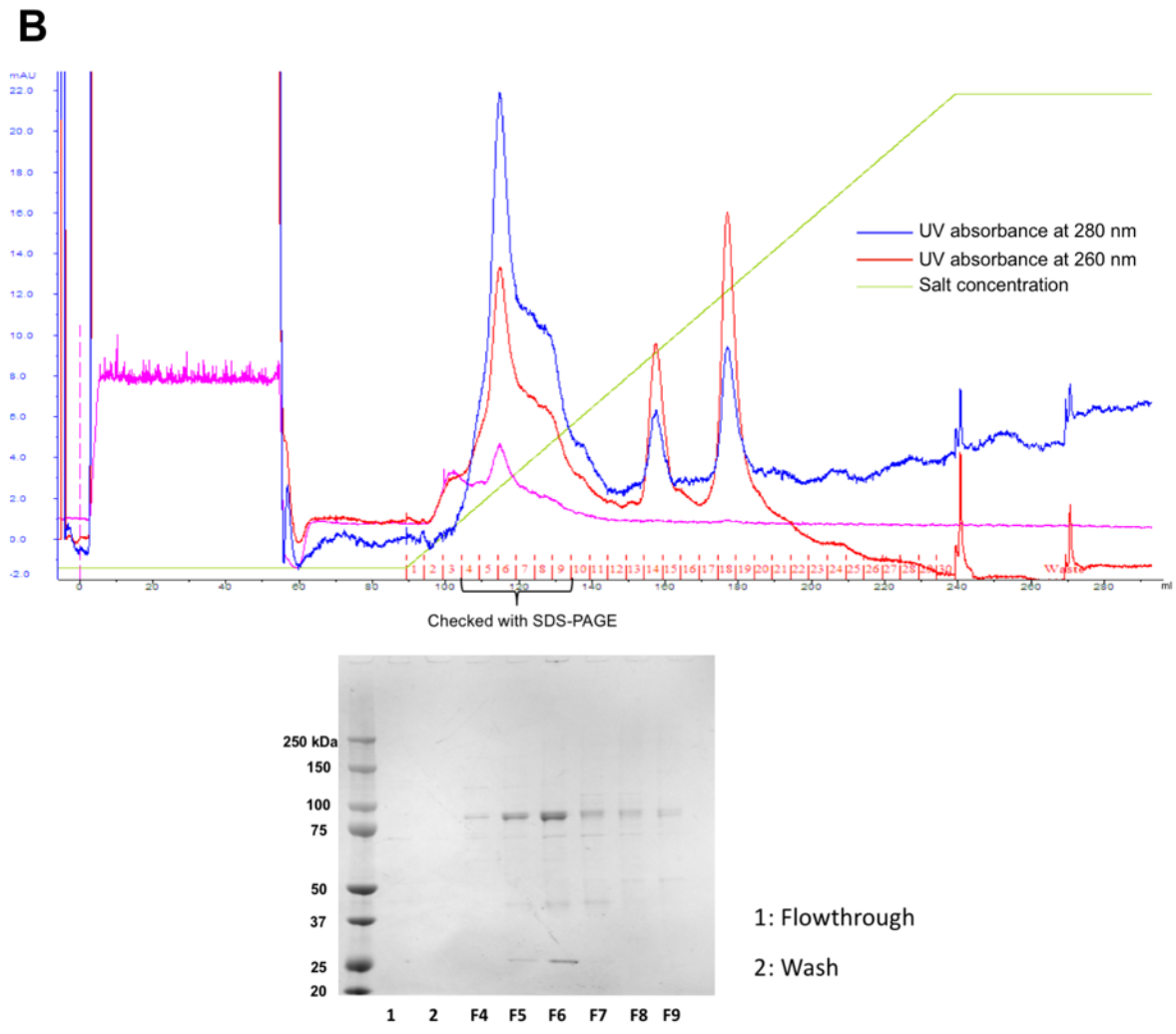
To start with, the Sf9 cells expressing Artemis, incubated for two days after the virus transformation, was collected for the purification and lysed in cell-lysis buffer (50 mM Tris pH 8.0, 500 mM KCl, 2 mM β-mercaptoethanol, 10% glycerol, 0.1% Triton X-100, and 20 mM imidazole supplemented with protease inhibitors). 10 ml of Ni-NTA resins (Qiagen) was used as the gravity column for the first-step His-tag purification (Figure 19A).

According to the gel, the expression of wild-type Artemis was successful using insect cells but a big part of the protein appeared to be insoluble and went into the pellet after the centrifugation. Nevertheless, the eluted fraction (50ml) was dialysed against 5L of low-salt buffer (50 mM Tris pH 8.0, 100 mM KCl, 2 mM β-mercaptoethanol, 10% glycerol, and 0.1% Triton X-100) for the Q-column purification. It was noted that white precipitate formed during dialysis. This could be Artemis aggregating as the protein may not be very stable under low-salt condition considering its flexibility. To avoid blocking the subsequent column, the dialysed fraction was filtered using the 0.2µm Sartorius™ Minisart™ Single use filter unit, loaded to the HiTrap Q column, washed and eluted with high-salt buffer of 1M KCl. Based on the ion-exchange chromatogram and the SDS-PAGE gel, the signal of UV 280nm was relatively low, and so was the signal of the band of wild-type Artemis (Figure 19B). This suggests that the precipitation during dialysis is very likely to be aggregated Artemis.

There were also bands and smearing of contamination of lower molecular weight. This could be caused by the degradation of Artemis compared to the elution fraction immediately after His-tag purification. Two bands around the molecular weight of 37 kDa and 25 kDa were later confirmed to be parts of Artemis by the Matrix-assisted laser desorption/ionization (MALDI) fingerprinting MS (PNAC facility, Department of Biochemistry). MALDI-MS showed that the two bands contained the peptides from different regions of Artemis although they are significantly smaller than the wild-type construct. It may be due to the low resolution of separation of different bands around those molecular weights or the band may be a mixture of different interacting peptides of Artemis.



Wild-type Artemis Nickel-affinity purification



Wild-type Artemis ion-exchange purification

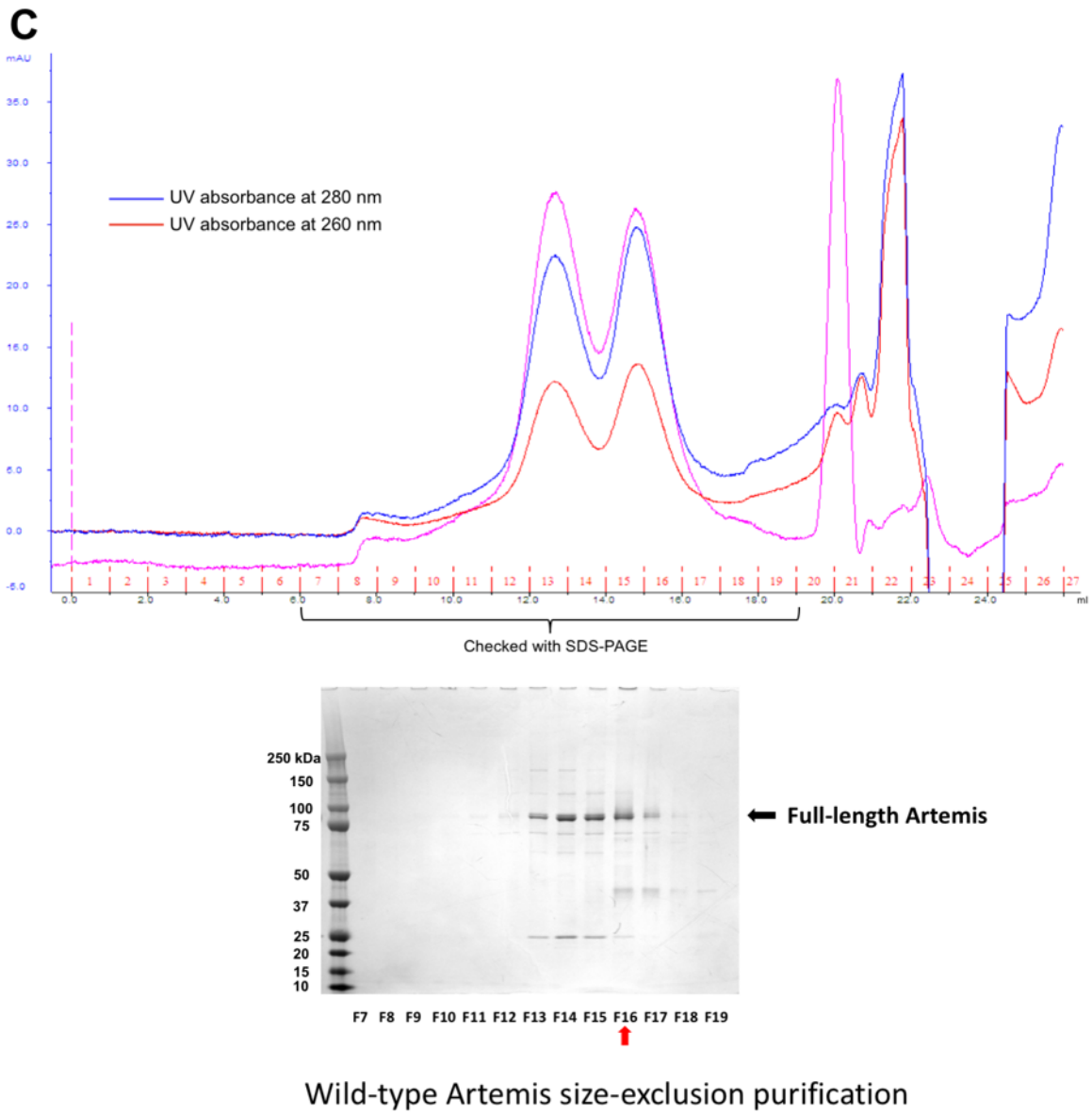


Figure 19. Purification of the wild-type Artemis. (A) SDS-PAGE gel of the His-tag purification of Artemis using Ni-NTA gravity flow; (B) The chromatogram and SDS-PAGE gel of the Q-column purification of Artemis using the HiTrap Q column; (C)The chromatogram and SDS-PAGE gel of the gel-filtration purification of Artemis using the Superose 6 10/300 column. Wild-type Artemis is highly flexible, degradable and heterogeneous. There is no canonical purification protocol and this purification was optimised based on the published ones. To confirm which fraction to collect, nuclease assay was used to check the final gel-filtration fractions and the nuclease assay will be introduced in detail in section 4.4. The numbers shown next to gel are the molecular weights (kDa) of the markers. F stands for Fraction in the lane labels, followed by the fraction number. The red up arrow identifies the final collected fraction. All SDS-PAGE gels were stained using Coomassie Blue.

Fractions containing the wild-type Artemis were than concentrated, filtered, centrifuged and loaded onto the Superose 6 10/300 column equilibrated with storage buffer (25 mM HEPES

pH 7.5, 500 mM KCl, and 2 mM DTT). During the process of centrifugation, the sample became more viscous with an increasing degree of yellow colour and with white precipitate appearing, indicating that the sample is easily saturated and not highly tolerant of the process of concentration. According to the Superose 6 10/300 chromatogram (Figure 19C), there are two main protein peaks at the elution volume of 13ml and 15ml, corresponding to the molecular weight of 669 kDa and 440 kDa based on the standard calibration curve from GE, which are much bigger than the size of wild-type Artemis (78 kDa). Meanwhile, the SDS-PAGE gel showed that the fractions of both peaks contained wild-type Artemis. As the standard proteins used for the calibration curve (Thyroglobulin for 669 kDa and Ferritin for 440 kDa) are structured and packed, the size of the Artemis may be at the same level considering that Artemis is highly flexible and unpacked, with the intrinsically disordered loops and tails extending from and floating around the structured core.

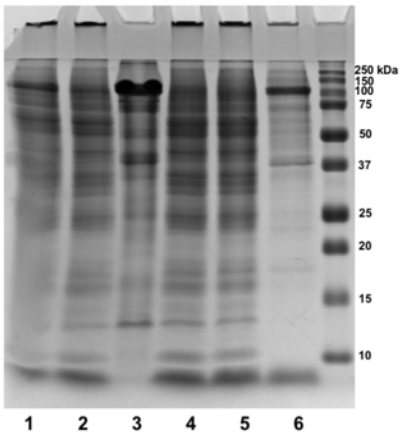
However, it was confusing as both peaks contained wild-type Artemis and it was unclear which peak should be collected. To ensure that the correct physiologic fraction of sample is collected, endonuclease assay was carried out on all the fractions from both peaks. A series of endonuclease activity assays were done and will be introduced systematically in chapter 5. The yield of the purification of wild-type Artemis is low. From two litres of insect cells, 100µg of the protein could be purified.

4.1.1.2 Purification of Artemis H115A

The purification of Artemis H115A is similar to that of the wild-type Artemis with the identical buffer conditions. However, this construct behaved a little differently. From the same amount of insect cell, there seemed to be more expression of Artemis H115A compared to the wildtype, along with more degradation. This was evident after the first step of His-tag purification (Figure 20A). Artemis H115A also precipitated during the dialysis with the low-salt buffer to go through Q-column purification. It was eluted at a similar level of salt during the Q-column elution step (Figure 20B). Due to the higher expression level, a larger volume of eluate was collected compared to that of the wildtype, resulting in a higher concentration level to load the sample onto the gel-filtration column (Superose 6 10/300). The Artemis H115A sample precipitated more than the wildtype during the concentrating process. The gel-filtration chromatogram is quite different to the one of wildtype (Figure 20C). There are two significant peaks in the chromatogram. One is in the void volume (around 7ml) and the other elutes at a similar position to the functional wild-type Artemis (around 15ml). According to the SDS-PAGE gel, there is continuous presence of the band of Artemis H115A from the void volume till the ideal elution position. This is likely to be caused by the over-centrifugation during the concentrating process of the sample preparation for the gel filtration. There was less sample during the purification of the wild-type Artemis and the centrifugation of the wild-type sample caused some small aggregates that ended up coming from the peak next to the functional batch of wild-type Artemis. While in this case of Artemis H115A, there was more protein expressed and the concentration increased compared to that of the wildtype, leading to various aggregates of different sizes ranging from the void volume to the region next to the ideal elution position.

To ensure that the correct fractions were collected, the chromatogram and SDS-PAGE gel were carefully compared to those of the wild-type Artemis and the fraction A15 was finally collected. The yield is higher than that of the wildtype. From two litres of insect cells, 300µg of the protein could be purified.

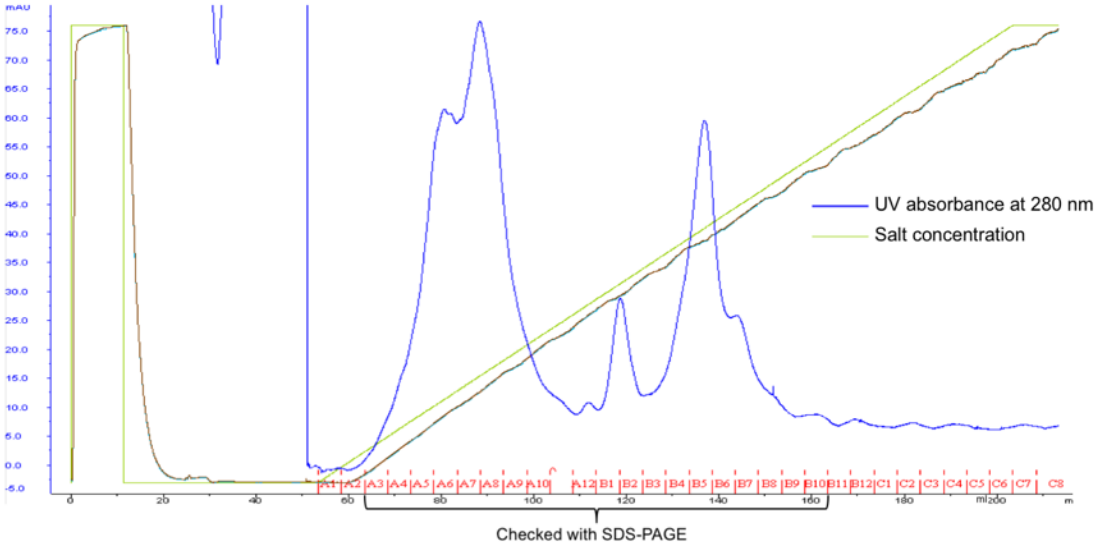
A

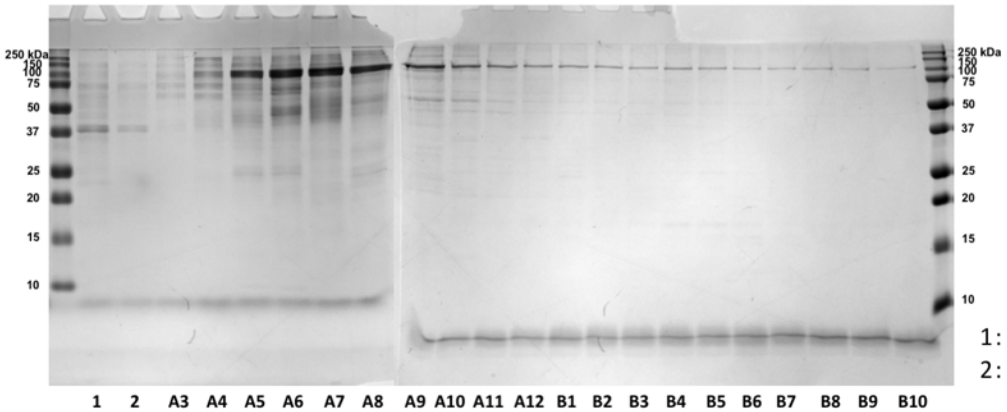


1 : Cell lysis
2 : Supernatant
3 : Pellet
4 : Flow Through
5 : Wash
6 : Elution

Artemis H115A Nickel-affinity purification

B





1: Flowthrough
2: Wash

Artemis H115A ion-exchange purification

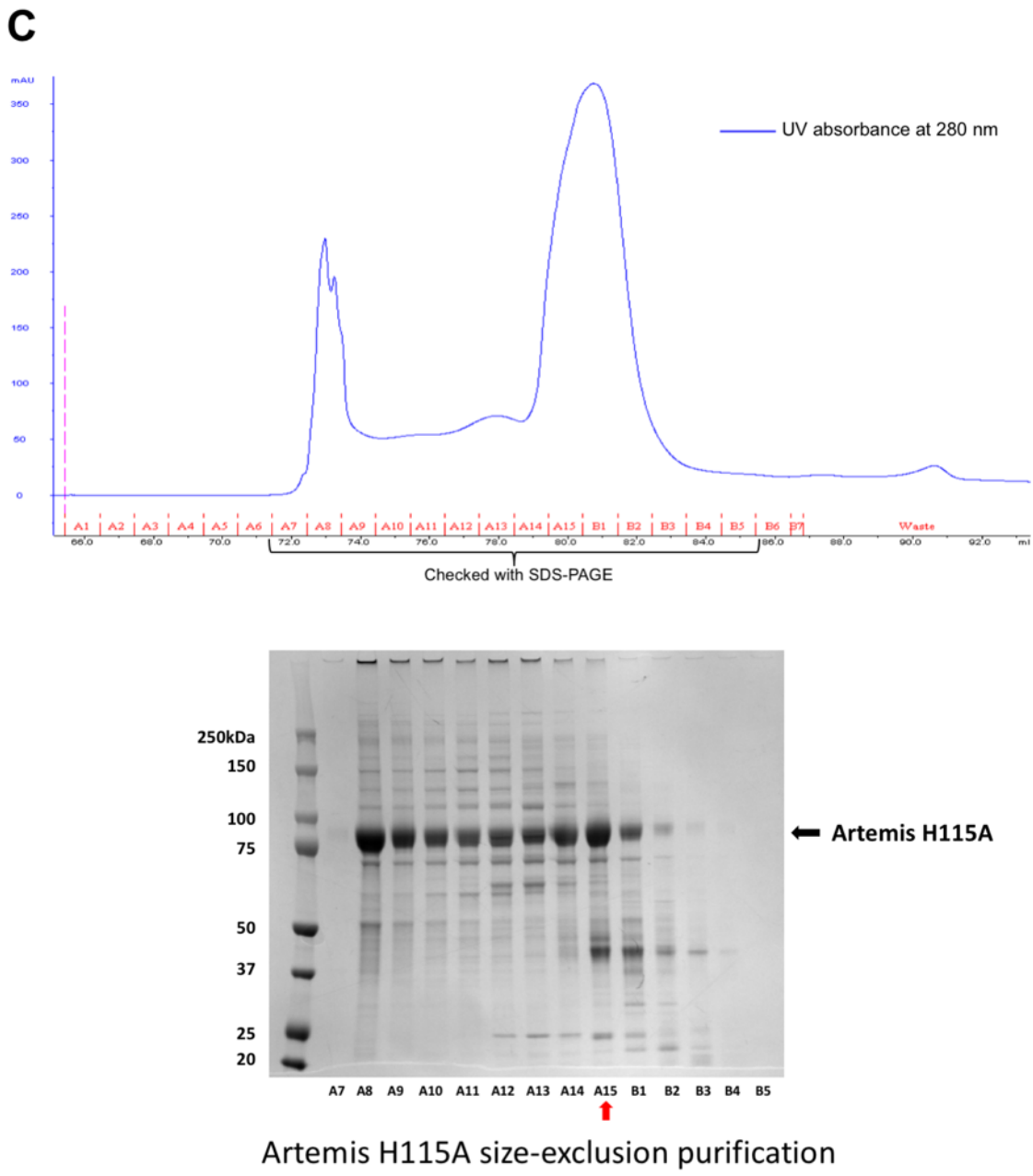


Figure 20. Purification of Artemis H115A. (A) His-tag purification of Artemis H115A using Ni-NTA; (B) Q-column purification of Artemis H115A using the HiTrap Q column; (C) Gel-filtration purification of Artemis H115A using the Superose 6 10/300 column. The purification of Artemis H115A is the same as that obtained from the optimised purification of wild-type Artemis. Compared to wild-type Artemis, Artemis H115A seems to have better expression while having decreased heterogeneity. To confirm which fraction to collect, the final gel-filtration fractions and chromatography were compared to those of wild-type Artemis. The numbers shown next to the gel are the molecular weights (kDa) of the markers. F stands for Fraction in the lane labels, followed by the fraction number. The red arrow indicates the final collected fraction. All SDS-PAGE gels were stained using Coomassie Blue.

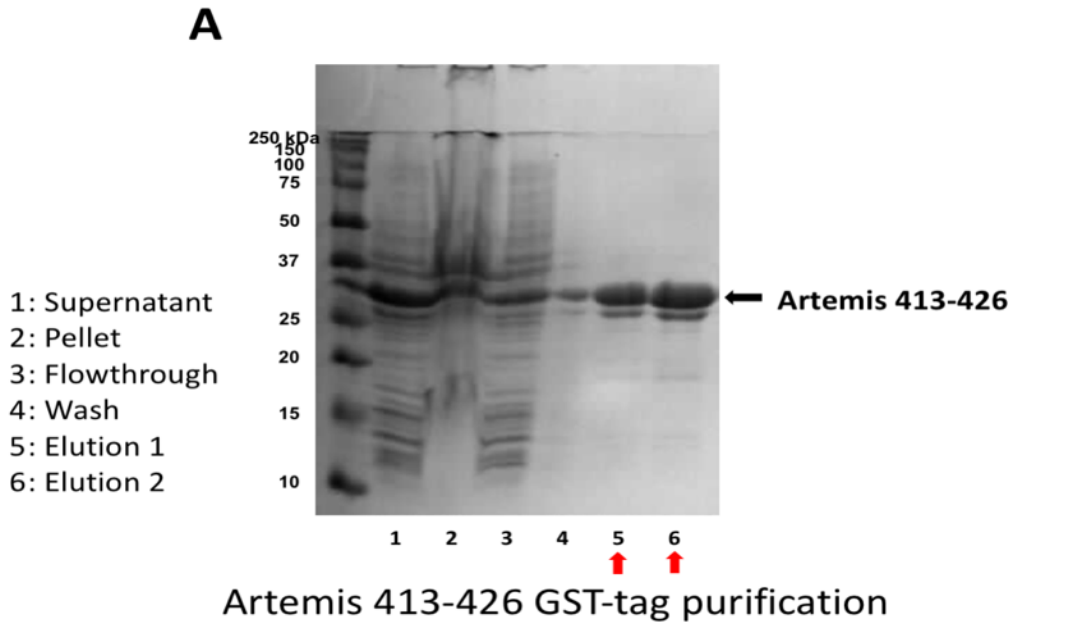
4.1.2 Purification of Artemis C-terminal Fragments

Artemis C-terminal fragments were produced to investigate the protein-protein interaction between DNA-PKcs and Artemis and identify the Artemis C-terminal region that binds to DNA-PKcs. Based on the bioinformatics analysis and previous research, five fragments were purified including Artemis 363-399, 399-408, 385-413, 399-426 and 413-426.

All the fragments are tagged with a N-terminal GST and His tag. The tag was designed to weight up the fragments, which are relatively small and difficult to purify on their own. Furthermore, it would be difficult to identify the small fragments without a GST tag on the SDS-PAGE gel during following experiments.

The purification of the fragments is a one-step GST-tag purification due to the high specificity of the GST tag and the low heterogeneity/ complexity of the fragments. For example, in the purification of Artemis 413-426 fragment, the transformed BL21 (DE3) cells was collected, lysed and centrifuged. The supernatant was loaded onto a GSTrap column, which was then washed and eluted with a high concentration of reduced L-glutathione. The elution fractions are pure with one major band of Artemis 413-426 and a small band at a lower position, which is likely to be the GST tag on its own as the fragment of 413-426 is unstructured and can be degraded easily (Figure 21A). The elution fractions then went through buffer exchange, concentrating and snap freezing, finally stored at -80 °C.

All the other fragments went through the same process and they were all pure as demonstrated in Figure 21B. They also all exhibit a small contamination band below the band of the purified fragment, suggesting that the contamination band is a common degraded product of all the fragment constructs, probably the GST tag.



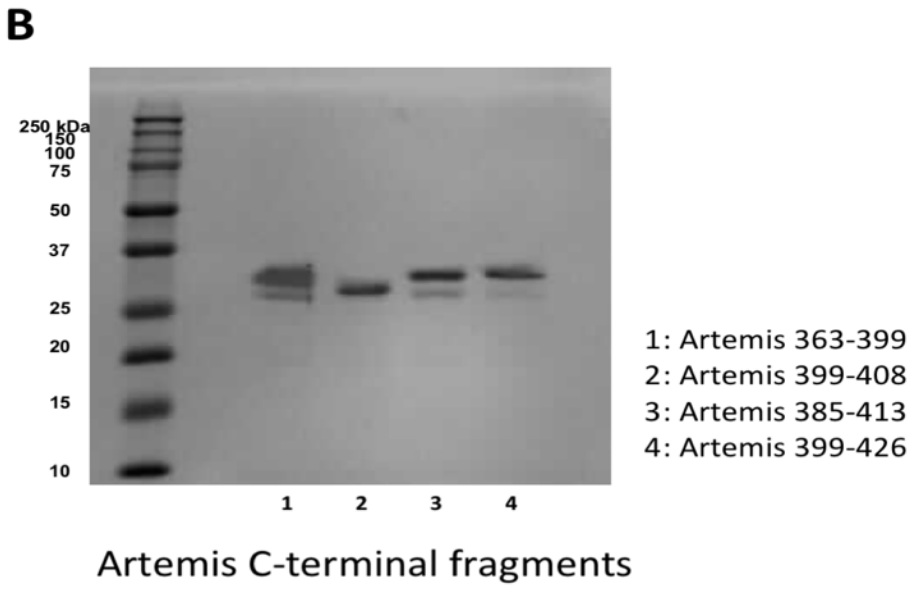


Figure 21. Purification of Artemis C-terminal fragments. (A) GST-tag purification of Artemis 413-426 using GSTrap column. The red arrow identifies the final collected fraction.; (B) Final purification results of the collected samples of other Artemis C-terminal fragments including Artemis 363-399; 399-408; 385-413; 399-426. The numbers shown next to the gels are the molecular weights (kDa) of the markers. All SDS-PAGE gels were stained using Coomassie Blue.

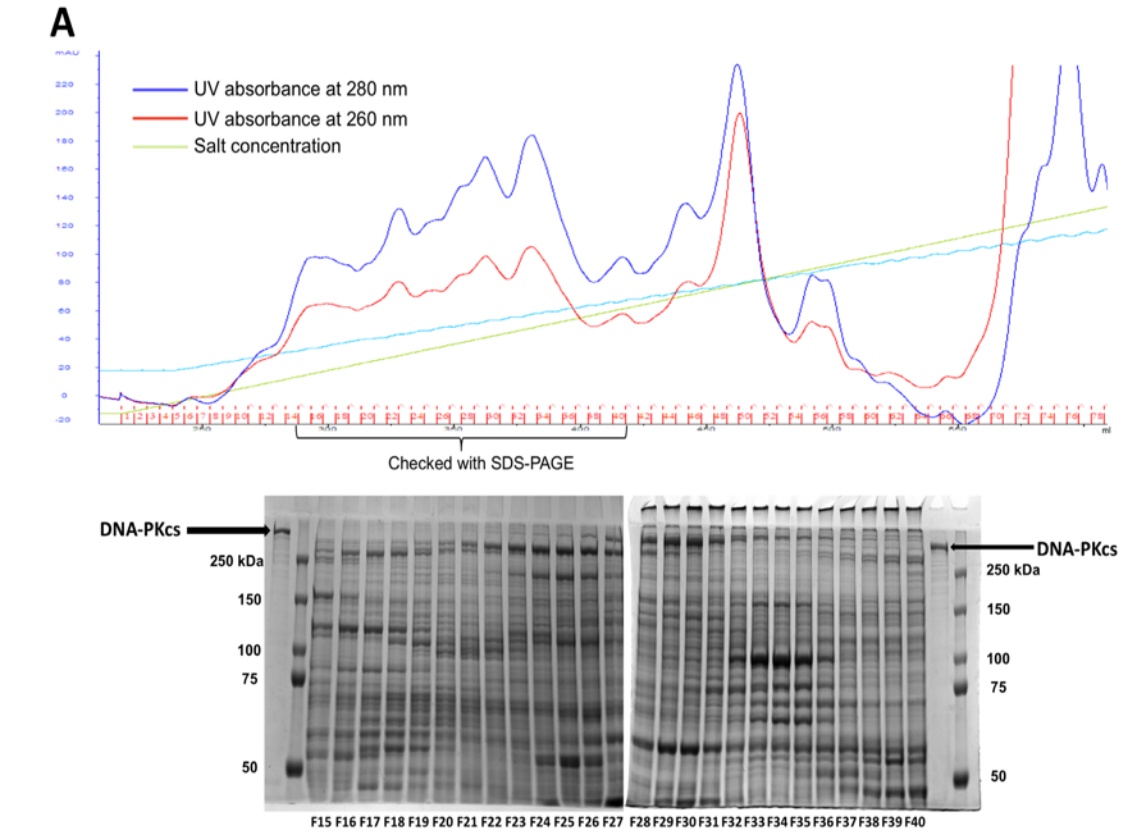
4.2 Purification of DNA-PKcs

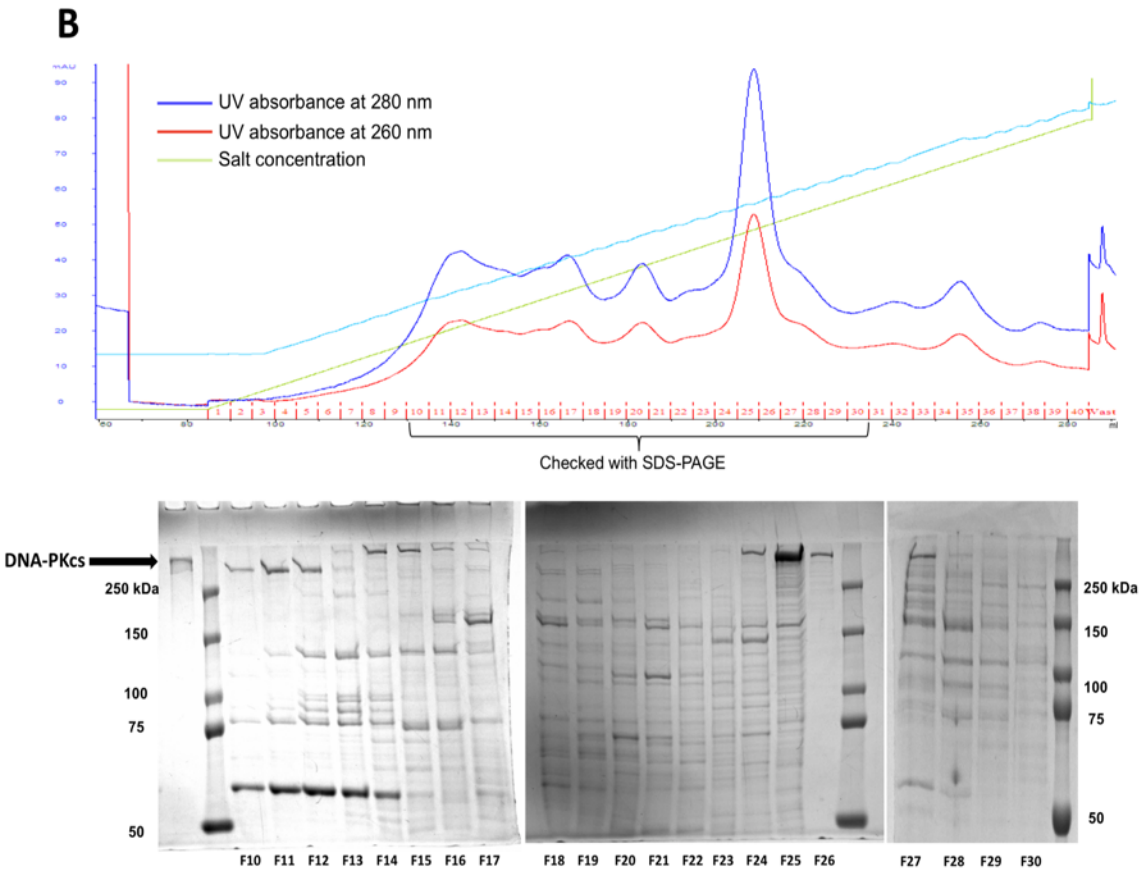
The purification of DNA-PKcs differs from that of the other recombinant constructs as DNA-PKcs is purified as a native protein. There is no affinity tag and the native expression level of DNA-PKcs in HeLa cells is less than that of the recombinantly expressed proteins. Careful identification of the correct band of DNA-PKcs is needed at the beginning to ensure that the right samples are collected for the following purification steps.

The material for the native purification of DNA-PKcs is HeLa cell nuclear extract, from which DNA-PKcs was actually first discovered in 1990 (Carter *et al.*, 1990). The purification of DNA-PKcs was optimised based on the published protocols (Gell and Jackson 1999; Sibanda *et al.,* 2017). There are four steps of purification in total: Q-column purification, heparin-column purification, S-column purification and gel-filtration purification.

Initially, Q column loading buffer (20 mM Hepes pH 7.6, 100 mM NaCl, 10%(v/v) glycerol, 0.5 mM EDTA, 2 mM MgCl2, 5 mM DTT) was added to thaw the frozen nuclear extract. The thawed mixture was then dialysed against the Q column loading buffer, loaded onto HiTrap Q column, washed and eluted from 100 mM NaCl to 500 mM NaCl. As shown in the SDS-PAGE gels (Figure 22A), pure DNA-PKcs sample, which was a gift from Dr Takashi Ochi, was used as a marker to help pick up the fractions containing DNA-PKcs. There was no significant peak in the chromatogram for the fractions containing DNA-PKcs and the fractions had complicated profiles of various protein bands compared to the classic affinity-tag purification of recombinant proteins.

The picked ones were then dialysed, loaded onto HiTrap heparin column, washed and eluted 100 mM NaCl to 1000 mM NaCl (Figure 22B). Compared to the picked fractions in the Q-column purification, the fractions with DNA-PKcs eluted as a significant peak in the chromatogram. Additionally, the SDS-PAGE gel profiles of the fractions became cleaner and the biggest contamination around 50 kDa was removed at lower-salt condition during the gradient elution.


A
UV absorbance at 280 nm
UV absorbance at 260 nm
Salt concentration
Checked with SDS-PAGE
DNA-PKcs
250 kDa
150
100
75
50
F15 F16 F17 F18 F19 F20 F21 F22 F23 F24 F25 F26 F27 F28 F29 F30 F31 F32 F33 F34 F35 F36 F37 F38 F39 F40


B
UV absorbance at 280 nm
UV absorbance at 260 nm
Salt concentration
Checked with SDS-PAGE
DNA-PKcs
250 kDa
150
100
75
50
F10 F11 F12 F13 F14 F15 F16 F17 F18 F19 F20 F21 F22 F23 F24 F25 F26 F27 F28 F29 F30

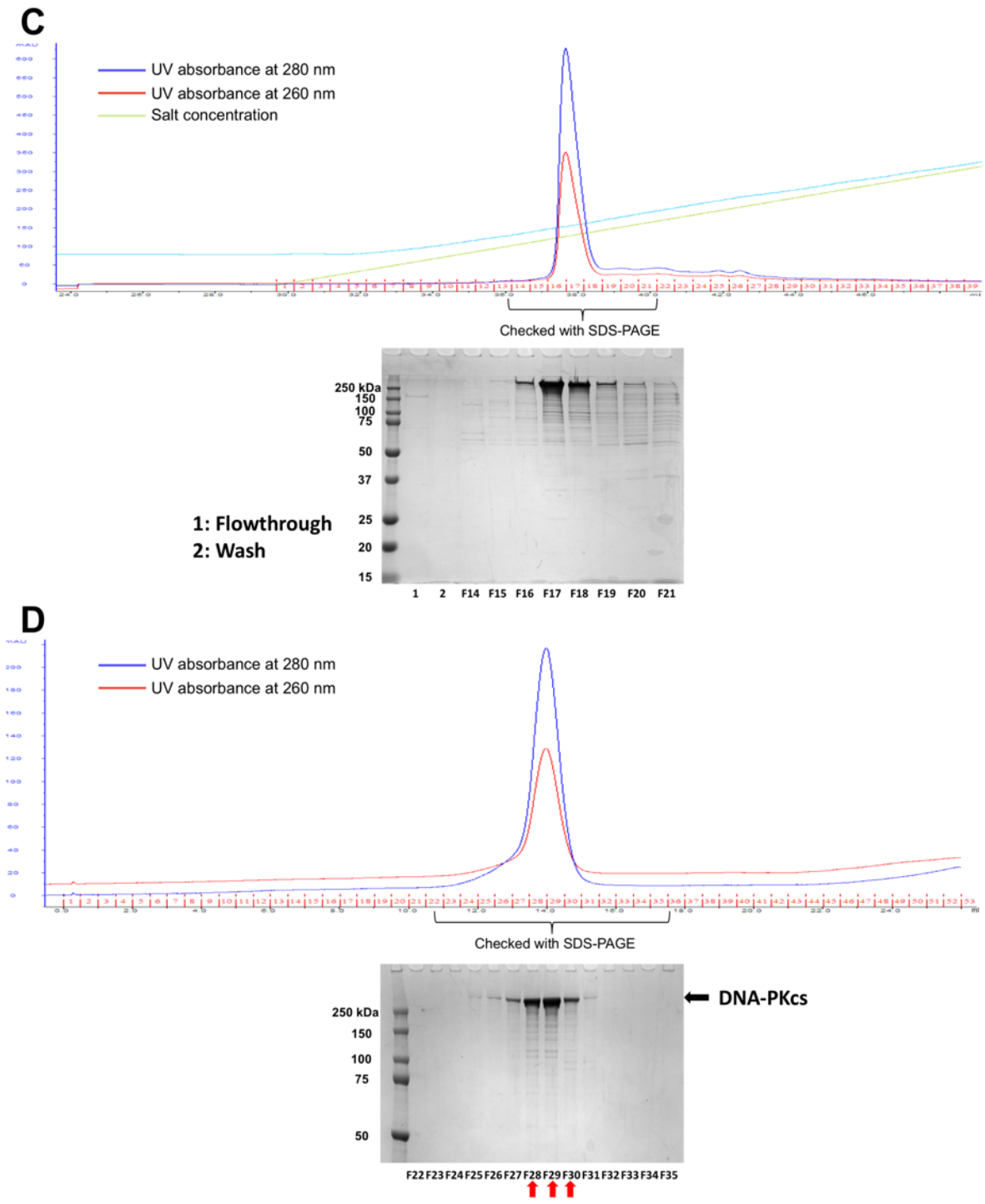


Figure 22. Purification of native DNA-PKcs. (A) Q-column purification of DNA-PKcs using HiTrap Q column; (B) Heparin-column purification of DNA-PKcs using the HiTrap Heparin column; (C) S-column purification of DNA-PKcs using Mono S column; (D) Gel-filtration purification of DNA-PKcs using Superose 6 10/300 column. The numbers shown next to gel are the molecular weights (kDa) of the markers. F stands for Fraction in the lane labels, followed by the fraction number. The red arrows identify the final collected fractions. All SDS-PAGE gels were stained using Coomassie Blue.

The following step was S-column purification using Mono S column (Figure 22C) (Buffers: 20 mM Hepes pH 7.6, 10%(v/v) glycerol, 0.5 mM EDTA, 2 mM MgCl2, 5 mM DTT with extra salt from 100 mM NaCl to 500 mM NaCl). This was to remove the degraded DNA-PKcs products. The fractions containing DNA-PKcs corresponding to the significant peak in the chromatogram were collected and concentrated for the final gel-filtration purification.

The chromatogram of the gel-filtration purification using Superose 6 10/300 column showed that the sample was pure with one peak only at the elution volume of 14 ml (Figure 22D) within the buffer (20 mM Hepes pH 7.6, 200 mM NaCl, 0.5 mM EDTA, 2 mM MgCl2, 5 mM DTT). The elution volume of DNA-PKcs is very close to that of Ferritin (440 kDa), which is of a similar size. The SDS-PAGE gel profile also showed clearly a strong band corresponding to DNA-PKcs, confirmed by mass spectrometry. There was some degradation of the protein but it was consistent with the published purified DNA-PKcs SDS-PAGE profiles as DNA-PKcs has a certain degree of flexibility.

4.3 Purification of DNA Ligase IV Constructs

The DNA ligase IV constructs purified include the DNA ligase IV along with the XRCC4 homodimer, the catalytic domain of DNA ligase IV and the DNA ligase IV DBD, of which only DNA ligase IV/ XRCC4 and DNA ligase IV DBD will be introduced here. The purification of the catalytic domain of DNA ligase IV is not described in this chapter but will be included in the supplementary information as the yield of the catalytic domain is too low for the subsequent protein biochemical and biophysical characterisation (Figure S1).

4.3.1 Purification of DNA Ligase IV Complex Constructs

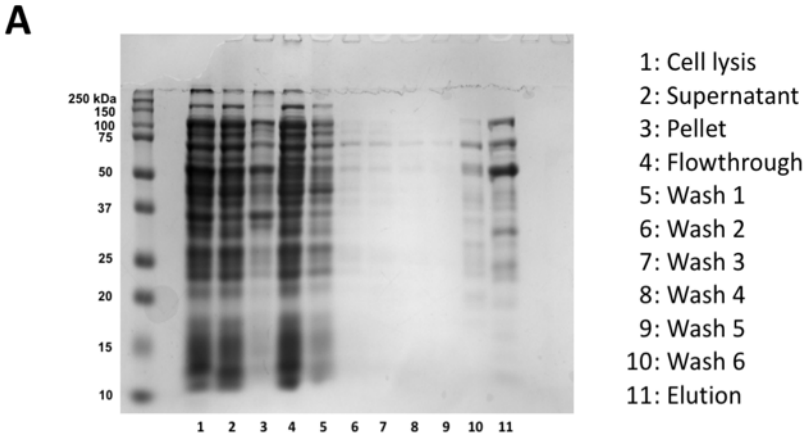
DNA ligase IV complex involves full-length DNA ligase IV and a XRCC4 homodimer. Due to the flexibility, the full-length DNA ligase IV was purified in complex with the XRCC4 homodimer that stabilises the C-terminal region. To express the complex, sequences of both DNA ligase IV and XRCC4 are put together on a co-expression vector (pRSFDuet1). Similar to Artemis, the His tag of DNA ligase IV is located on the C terminus to cope with the flexibility of C-terminal region of DNA ligase IV. Two constructs of the complex were purified: the wild-type DNA ligase IV complex and the mutant DNA ligase IV complex (LigIV K273A), which is non-functional. The purification protocol is modified from that published by Wang et al. in 2007. and involves four steps including Nickel-affinity purification, ion-exchange purification, gel-filtration purification and hydroxyapatite purification. The purification of the DNA ligase IV complex is similar to that of full-length Artemis due to heterogeneity of the protein samples. It was challenging but the purification protocol succeeded in removing most of the heterogeneity as shown by the SDS-PAGE gel.

4.3.1.1 Purification of Wild-Type DNA Ligase IV Complex

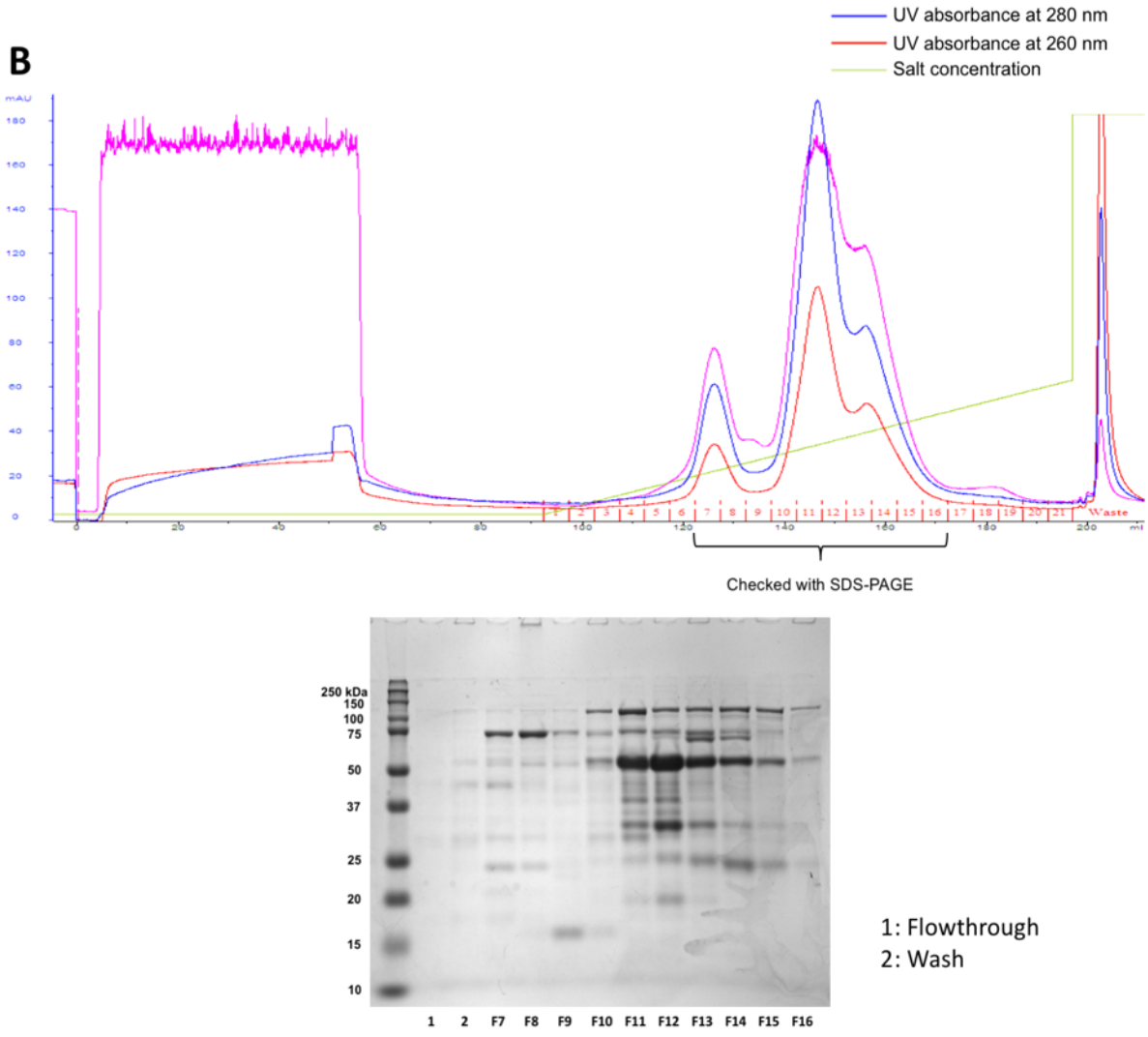
The Rosetta cells carrying the DNA ligase IV/ XRCC4 co-expression vector, induced with IPTG, were incubated at 16 °C and 220 rpm overnight. After incubation, the cells were collected, lysed and centrifuged (Buffer: 50 mM sodium phosphate buffer pH 8.0, 300 mM NaCl, 10%(v/v) glycerol, 1 mM β-mercaptoethanol, 5 mM imidazole, 100 uM PMSF, 1x Complete protease inhibitor EDTA-free). The supernatant was applied to the first step of purification using HisTrap column. For the washing step, six different buffers were explored with changing salt or imidazole concentrations. According to the SDS-PAGE gel (Figure 23A), there were three main bands. The one over 100 kDa is full-length DNA ligase IV, the one lower than 75 kDa is the degraded DNA ligase IV and the one around 50 kDa is full-length XRC44, confirmed by mass spectrometry. It should be noted that the actual molecular weight of XRCC4 is 38kDa, which should appear close to the band of 37 kDa in the marker. However, this phenomenon is consistent with the previous purification profiles published (Lee et al., 2000; Wang et al., 2007) and it is likely to be caused by the amino acid composition of XRCC4.

The elution fraction was then adjusted to lower salt concentration (100 mM) with a no-salt buffer to be loaded onto HiTrap Q column for the second-step purification. During the Q-column purification, some degraded DNA ligase IV and contaminations with low molecular weight were removed (Figure 23B). The fractions containing DNA ligase IV and XRCC4 were collected and concentrated for the gel-filtration purification.

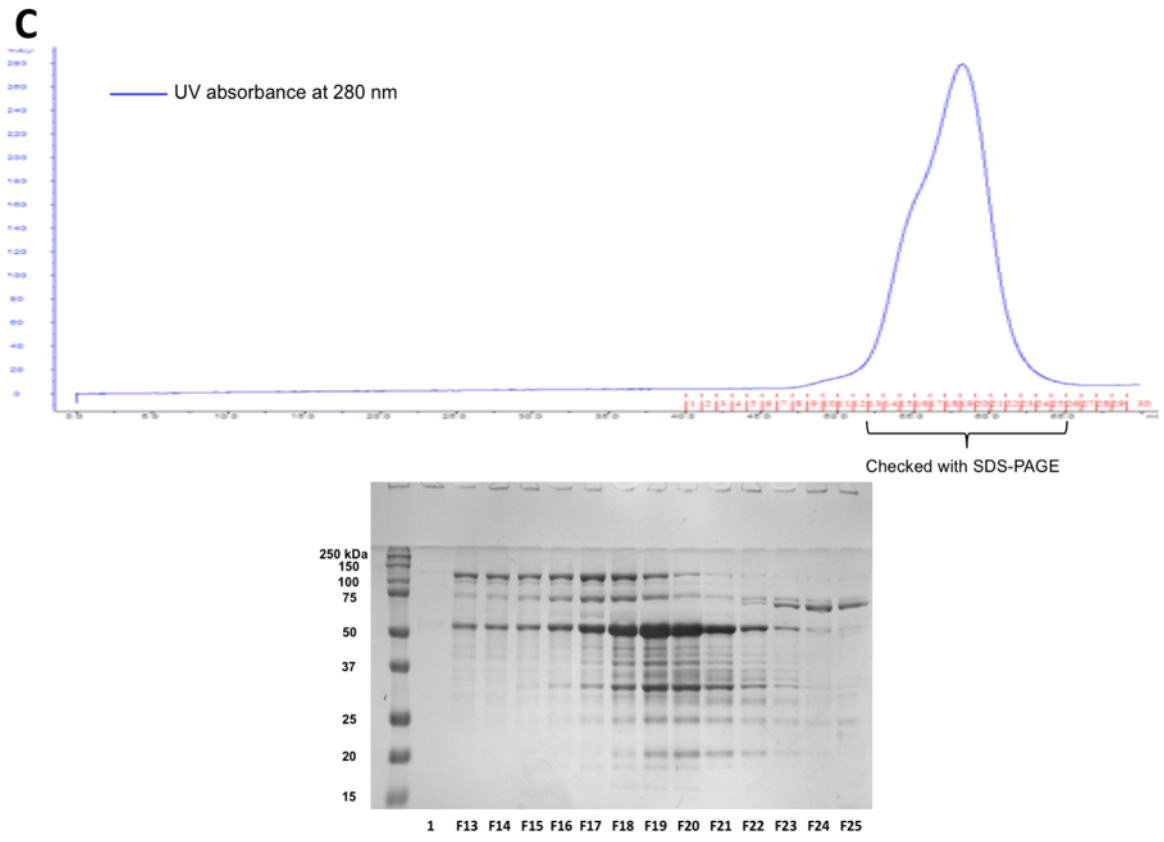
Superdex 200 16/60 was used for gel filtration (Figure 23C). Most of the excess XRCC4 and degraded products of lower molecular weight were removed during this step as they eluted at the position of 60 ml. Although the heterogeneous sample resulted in broad and continuous peaks without clear separation, the fractions containing most of the DNA ligase IV complex appeared at the position of 55 ml. It is close to the elution position of ferritin (440 kDa), confirming the flexibility of DNA ligase IV complex which is likely to be caused by the flexible C-terminal region of DNA ligase IV. The fractions with DNA ligase IV complex and no excess of XRCC4 were collected and dialysed for hydroxyapatite purification.



DNA ligase IV complex Nickel-affinity purification



DNA ligase IV complex ion-exchange purification



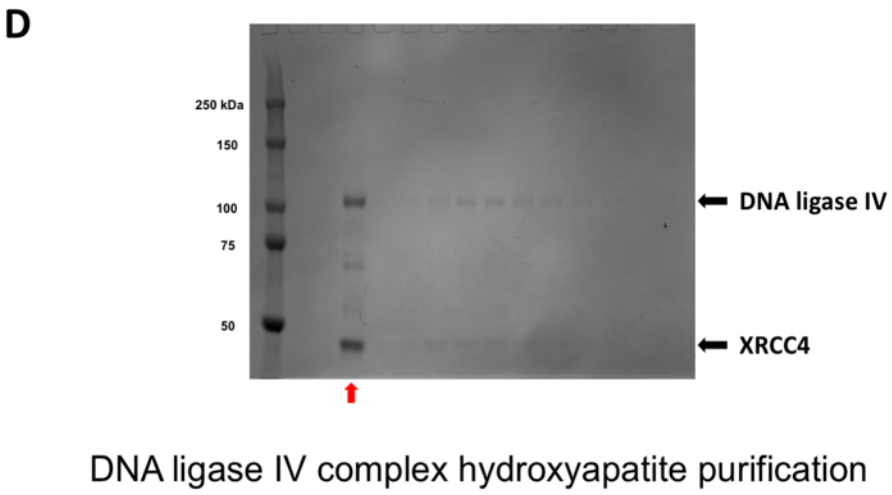


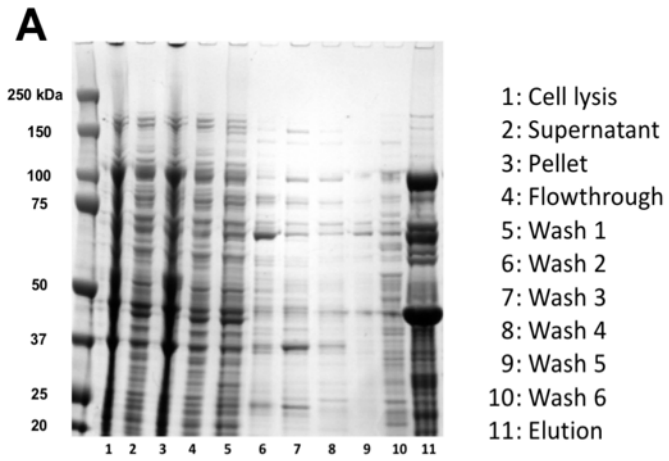
Figure 23. Purification of wild-type DNA ligase IV complex. (A) His-tag purification of wild-type DNA ligase IV complex using HisTrap column; (B) Q-column purification of wild-type DNA ligase IV complex using the HiTrap Q column; (C) Gel-filtration purification of wild-type DNA ligase IV complex using superdex 200 16/60 column; (D) Hydroxyapatite purification of wild-type DNA ligase IV complex using Bio-Scale CHT Type I column. The numbers shown next to gel is the molecular weight (kDa) of the markers. F stands for Fraction in the lane labels, followed by the fraction number. The red up arrow is pointing the final collected fraction. All SDS-PAGE gels were stained using Coomassie Blue.

Bio-Scale CHT Type I column, used for the final hydroxyapatite purification, was the final step in the removal of nuclease contaminants (Wang et al., 2007) (Figure 23D). The final fraction collected was relatively pure with a little degraded DNA ligase IV around 75 kDa. The yield of the protein from a six-litre culture was 2 mg.

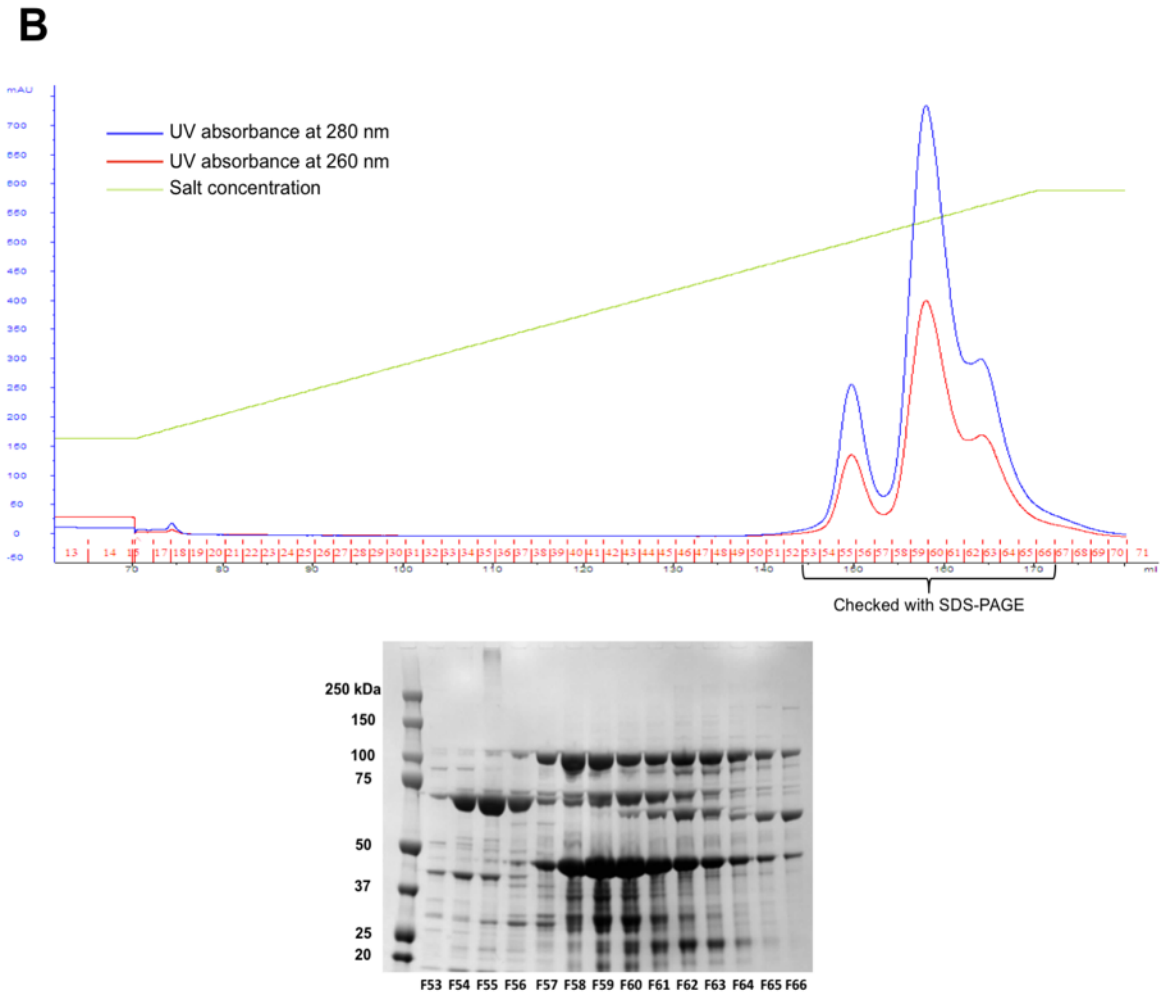
4.3.1.2 Purification of Mutant DNA Ligase IV (LigIV K273A) Complex

K273 of DNA ligase IV reacts with ATP to form the lysine-AMP covalent intermediate at the beginning of DNA ligation. The mutation (K273A) inhibits the process of DNA ligase IV adenylation, making the mutant DNA ligase IV (LigIV K273A) complex non-fuctional.

The purification of the mutant complex is the same as that of the wildtype. However, the mutant complex behaved slightly differently. From the same volume of cells, more mutant protein complex than wildtype complex is obtained after His-tag purification but with more contamination bands (Figure 24A), indicating that the mutant complex has better expression or can be more soluble. Moreover, in the step of Q-column purification, although the elution peaks in the chromatogram are similar in shape, those from the mutant purification are at a higher salt concentration level (Figure 24B). This suggests that the wildtype and mutant should have different ionic properties and the mutant complex should have more exposed negative-charge surface area in the buffer. During the gel-filtration purification, the peak of the mutant eluted later after 60 ml (Figure 24C). Furthermore, unlike the wildtype, the peaks of the mutant complex and XRCC4 and other contaminants are all merged into one. This showed that the molecule of the mutant complex may be more compact than the wild-type molecule. Last but not least, the hydroxyapatite purification was the last step and the final fractions were collected (Figure 24D). The yield of the mutant complex is better than the wildtype. About 4 mg of protein was finally collected from a six-litre culture.



Mutant DNA ligase IV complex (LigIV K273A) Nickel-affinity purification



Mutant DNA ligase IV complex (LigIV K273A) ion-exchange purification

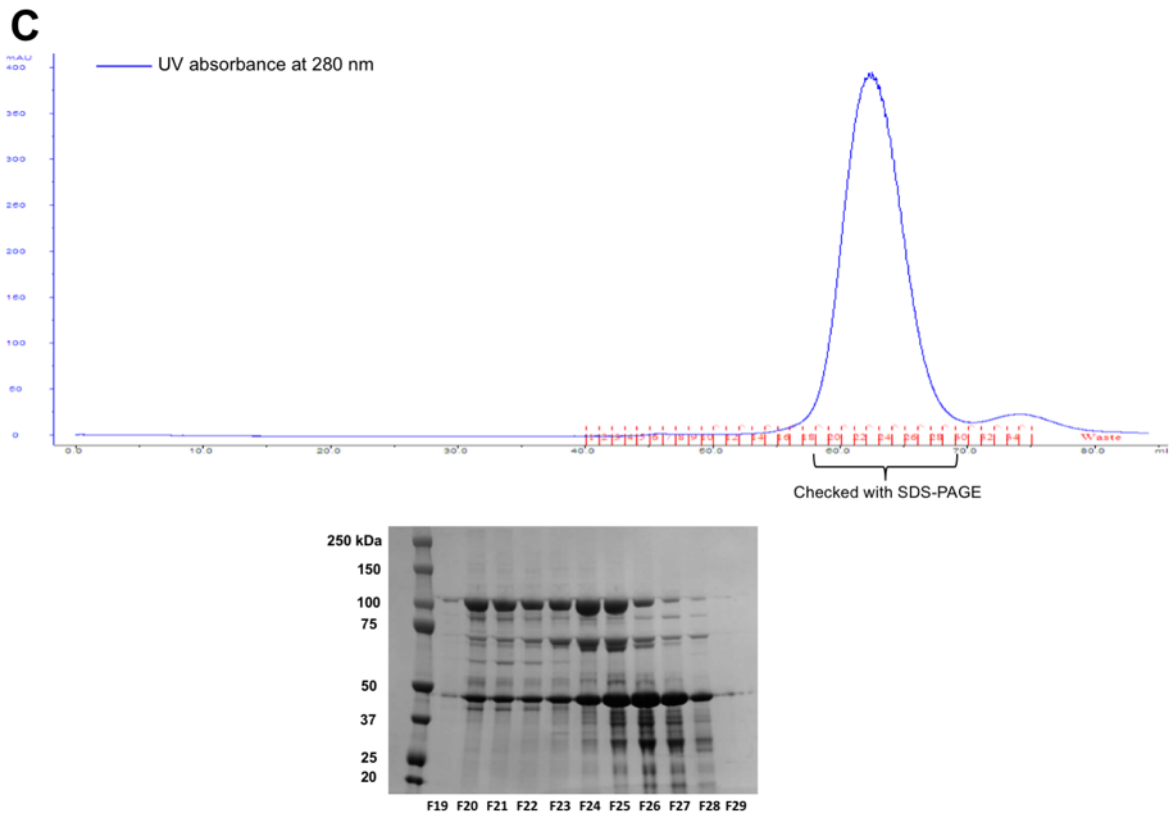


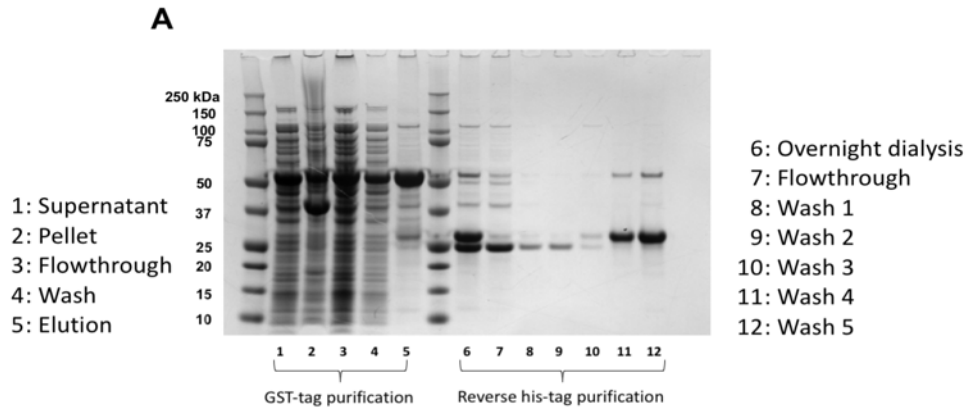
Figure 24. Purification of mutant DNA ligase IV complex (LigIV K273A). (A) His-tag purification of mutant DNA ligase IV complex using HisTrap column; (B) Q-column purification of mutant DNA ligase IV complex using the HiTrap Q column; (C) Gel-filtration purification of mutant DNA ligase IV complex using Superdex 200 16/60 column; (D) Hydroxyapatite purification of mutant DNA ligase IV complex using Bio-Scale CHT Type I column. The numbers shown next to the gel are the molecular weights (kDa) of the markers. F stands for Fraction in the lane labels, followed by the fraction number. The red arrow indicates the final collected fraction. All SDS-PAGE gels were stained using Coomassie Blue.

4.3.2 Purification of DNA Ligase IV DNA Binding Domain

The DNA ligase IV DBD construct comprises residues 1-230. It was purified for the fragment-based drug discovery targeting the DNA ligase IV-Artemis interaction site. The sequence of DNA ligase IV DBD was inserted in the pGAT3 vector with a N-terminal His-GST tag. The purification process includes three steps: GST-tag purification, reverse His-tag purification and gel-filtration purification.

The BL21 (DE3) cells expressing DNA ligase IV DBD were collected, lysed and centrifuged with the supernatant loaded onto the GSTrap column (Lysis buffer: 50 mM Tris-HCl pH 8, 300 mM NaCl, 5 mM DTT, 1x Complete protease inhibitor EDTA-free). The column was then washed and eluted. The elution profile showed a strong band of the DNA ligase IV DBD with the His-GST tag slightly above the band of 50 kDa in the marker (Figure 25A). The eluted sample was then added with TEV protease and dialysed at 4 °C overnight (Dialysis buffer: 30 mM Tris-HCl pH 8, 150 mM NaCl, 2 mM β-mercaptoethanol). Later, the dialysed sample was loaded onto HisTrap column preequilibrated with the dialysis buffer and the flowthrough collected during the reverse His-tag purification. This is to remove the undigested protein, cleaved His-GST tag and the His-tagged TEV proteases. The column was then washed using buffers with increasing imidazole concentrations. According to the gel, there were tag-cleaved DNA ligase IV DBD samples non-specifically binding to the column but they were washed off in the first two runs of washing with low-imidazole buffer (Figure 25A).

The flowthrough and wash fractions containing DNA ligase IV DBD were then concentrated and loaded onto the Superdex 75 16/60 column (Final buffer: 25 mM Tris-HCl pH 7.5, 150 mM NaCl, 5 mM DTT). According to the gel-filtration chromatogram and SDS-PAGE gel, the uncleaved protein, GST tag and other contaminants of higher molecular weight were all removed before the DNA ligase IV DBD came out at the corresponding position (Figure 25B). The peak fractions were then concentrated and collected. From four litres of cells, over 10 mg of the protein was purified.



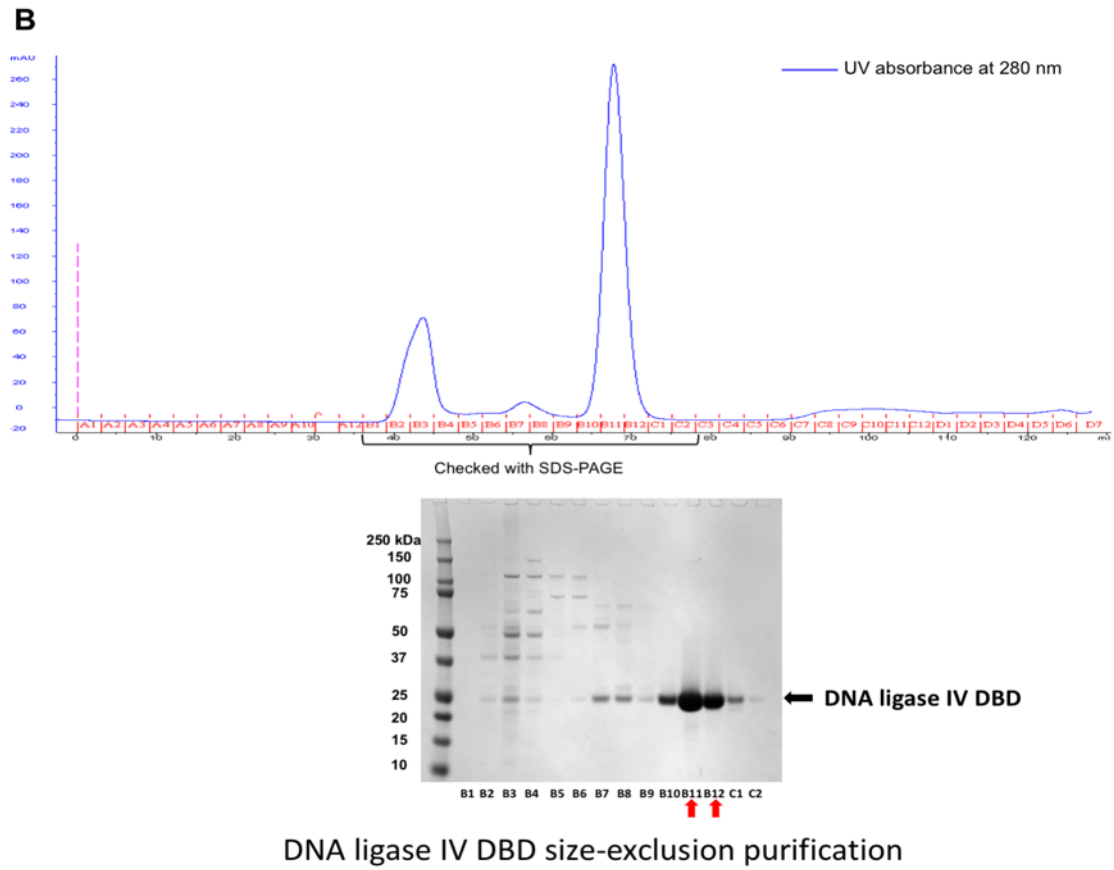
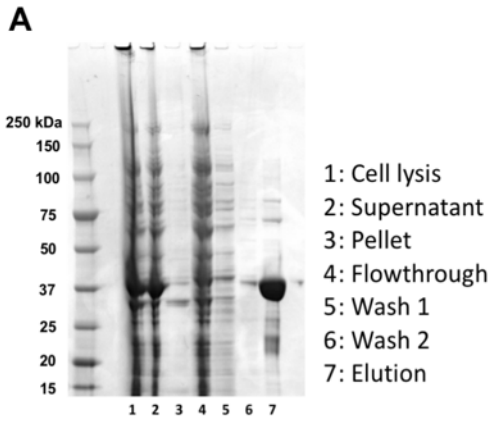


Figure 25. Purification of DNA ligase IV DBD (DNA ligase IV 1-230). (A) GST-tag purification and reverse His-tag purification of DNA ligase IV DBD using HisTrap column; (B) Gel-filtration purification of DNA ligase IV DBD using the Superdex 75 16/60 column. The numbers shown next to the gel are the molecular weights (kDa) of the markers. The red arrow indicates the final collected fraction. All SDS-PAGE gels were stained using Coomassie Blue.

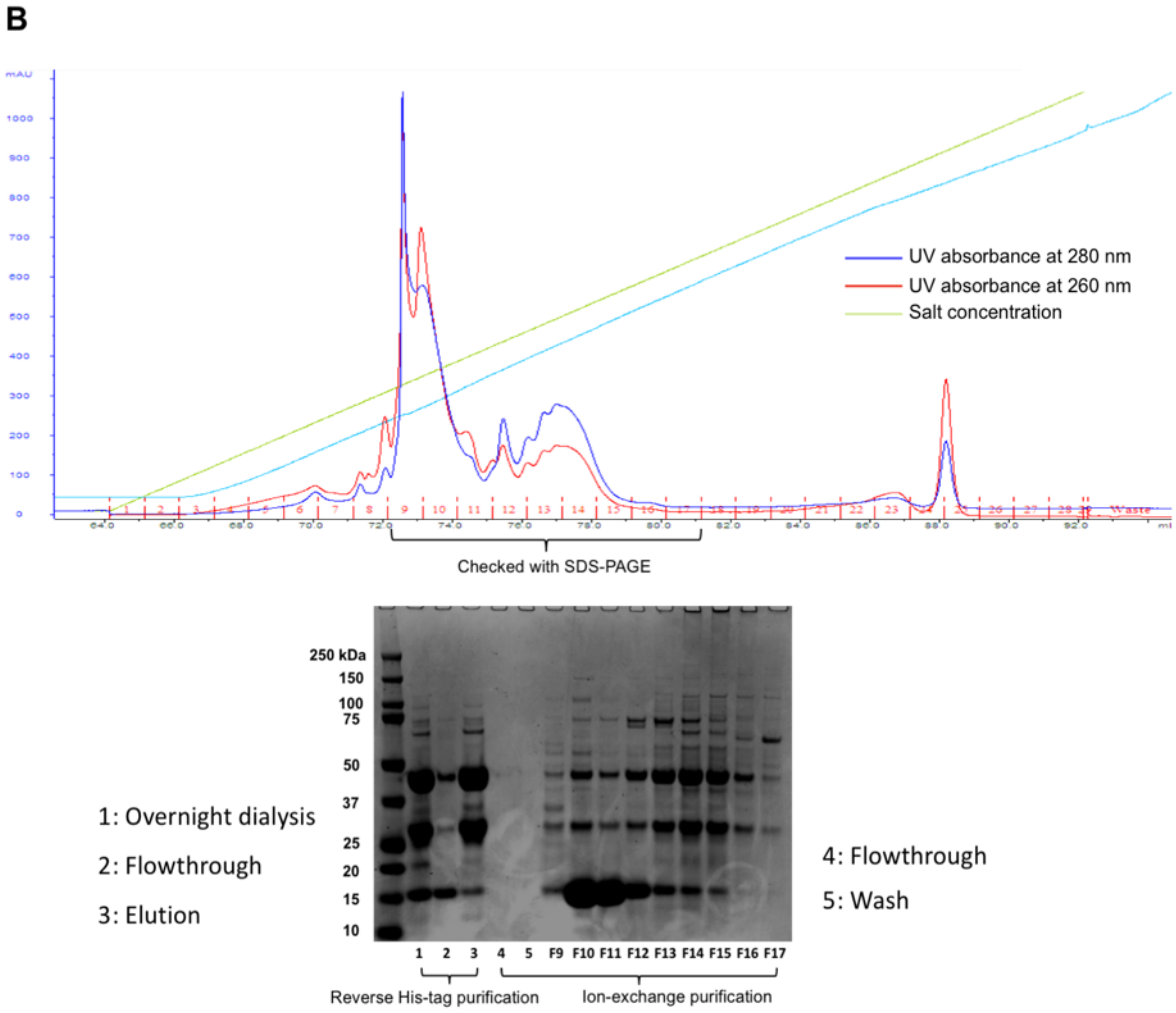
4.4 Purification of Ku80 CTD

Ku80 593-732 corresponds to the C-terminal globular domain of Ku80 and the DNA-PKcs interaction sequence. The sequence of Ku80 CTD was inserted in the pGAT3 vector with a N-terminal His-GST tag. The purification of Ku80 CTD is based on the published method (Sibanda *et al.,* 2010). There are five steps in total including His-tag purification, reverse His-tag purification, Q-column purification, reverse GST-tag purification and gel-filtration purification.

The BL21 (DE3) cells expressing Ku80 CTD were collected, lysed and centrifuged with the supernatant loaded onto Ni-NTA beads (Lysis buffer: 20 mM Tris pH 7.5, 150 mM NaCl, 2mM β-mercaptoethanol, 1x Complete protease inhibitor EDTA-free). The beads were then washed twice with low-imidazole buffer and eluted with high-imidazole buffer. There was a strong band of protein around 37 kDa in the elution that should correspond to the GST-tagged Ku80 CTD (Figure 26A). TEV protease was added to the eluted sample and dialysed at 4 °C overnight (20 mM Tris pH 7.5, 10 mM NaCl, 2 mM β-mercaptoethanol). During the following reverse His-tag purification, the dialysed sample was loaded on to HisTrap column with the flowthrough collected to remove the uncut protein and the cleaved GST tags. The TEV protease was not highly active in this case and a large proportion of the sample remained uncut (Figure 26B). The flowthrough was directly loaded onto the mono Q column for purification with an elution salt gradient from 10 mM NaCl to 1 M NaCl. According to the chromatogram and SDS-PAGE gel, this step removed most of the contaminants of higher molecular weight including the GST tag and the uncut protein. The fractions containing most Ku80 CTD were collected. However, to remove remaining GST tag and uncut protein, the collected fractions were loaded on to GSTrap column for reverse GST-tag purification. The flowthrough was collected and the GST and uncut protein were mostly removed (Figure 26C). The flowthrough was subsequently concentrated and loaded onto Superdex 200 10/300 (Final buffer: 20 mM HEPES pH 8.0, 150 mM NaCl, 5 mM DTT). Most of the contaminants were removed and the final fractions containing Ku80 CTD were collected and concentrated for storage (Figure 26C). From four litres of cells, 5 mg of Ku80 CTD was purified.



Ku80 CTD Nickel-affinity purification



Ku80 CTD reverse His-tag purification and ion-exchange purification

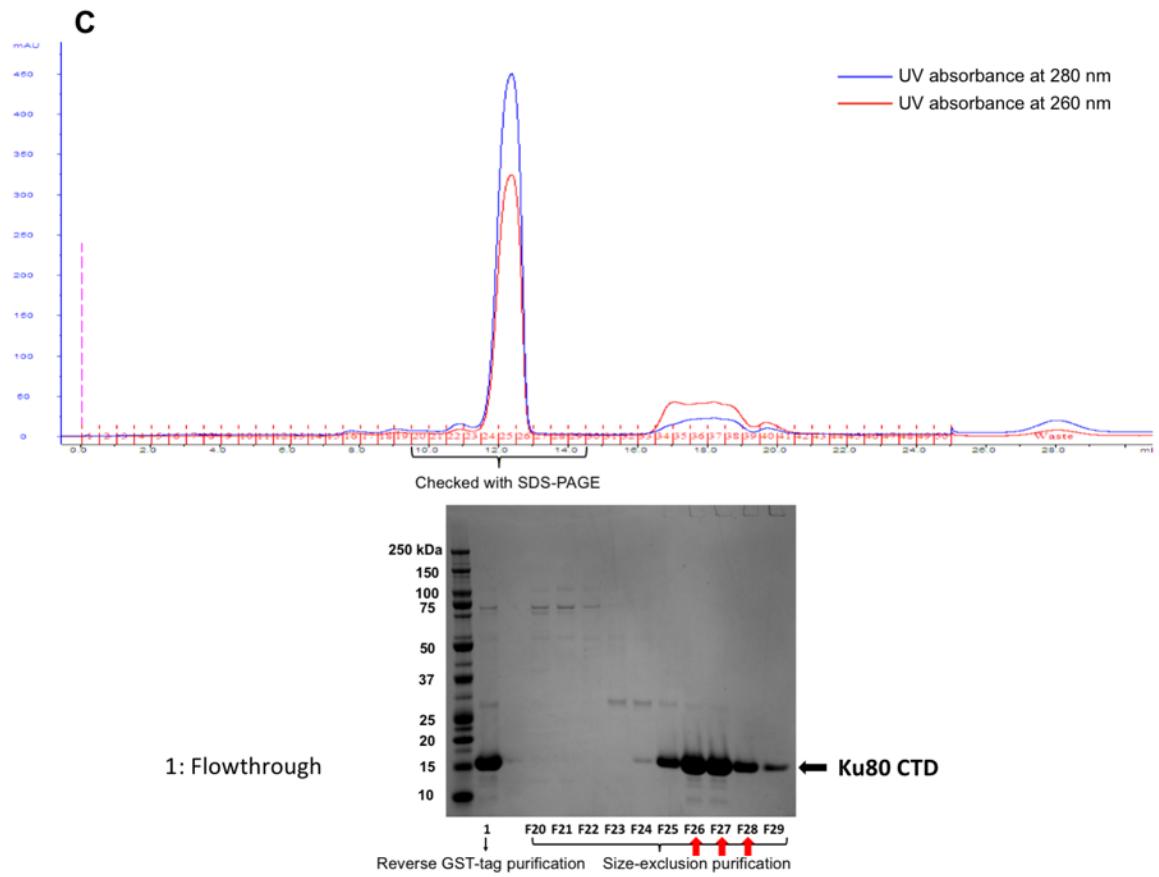


Figure 26. Purification of Ku80 CTD (Ku80 593-732). (A) His-tag purification of Ku80 CTD using Ni-NTA; (B) Reverse His-tag purification and Q-column purification of DNA ligase IV DBD using HisTrap and mono Q column; (C) Reverse GST-tag purification and gel-filtration using GSTrap and Superdex 200 10/300 column. The numbers shown next to gel are the molecular weights (kDa) of the markers. F stands for Fraction in the lane labels, followed by the fraction number. The red arrow indicates the final collected fraction. All SDS-PAGE gels were stained using Coomassie Blue.

4.5 Summary

The purifications of different NHEJ constructs presented various challenges. Many of them are highly flexible, leading to high level of heterogeneity and degradation/contamination and multiple steps of purifications. Good examples include the purifications of full-length Artemis constructs and DNA ligase IV complexes. Moreover, the yields from the purifications of many human NHEJ proteins were relatively low, including those of full-length Artemis and DNA ligase IV complexes, not to mention the purification of native DNA-PKcs.

Nevertheless, after purification and optimisation, most of the contamination was removed and the pure samples were successfully collected. In addition, although limited by the low yields of many proteins, biochemical and biophysical characterisations proved useful for investigating the properties of the proteins and NHEJ, which will be discussed in chapter 5. Last but not least, the purified samples were used for the later cryo-EM structure studies, which will be introduced in chapter 6.

Chapter 5. Protein Biochemical and Biophysical Characterisation

The purified proteins were used for biochemical and biophysical characterisation to study a series of questions relevant to targeting the DNA-PKcs/Artemis endonuclease complex, including its activity and interactions with other protein of NHEJ. Moreover, various purified constructs were used for the collaboration between our group and the Strick group to study the temporal organisation of NHEJ.

The first priority was to understand the structure and function of the DNA-PKcs/ Artemis endonuclease complex. To investigate this, purified Artemis and DNA-PKcs proteins were used in a series of nuclease assays with different NHEJ components. The Artemis H115A mutant was used in preliminary structural studies. Pulldown experiments were conducted to study the protein-protein interactions between DNA-PKcs and Artemis and to define the Artemis region interacting with DNA-PKcs.

Secondly, preliminary screening experiments were initiated targeting the interaction site between Artemis and DNA ligase IV with a view to initiating a fragment-based drug discovery campaign.

Last but not least, experiments were conducted to understand the temporal organisation of NHEJ, especially the end synapsis. These required purified constructs of DNA-PKcs, DNA ligase IV complex, and PAXX protein for the collaboration work using single-molecule methods with the group of Terence Strick in the Institut Jacques Monod in Paris.

5.1 DNA-PKcs/Artemis Characterisation

Functional assay was conducted to understand the activity of DNA-PKcs/Artemis endonuclease complex. Biophysical experiments were used to characterise Artemis H115A. Different Artemis C-terminal peptides were applied to pulldown experiments with DNA-PKcs.

5.1.1 DNA-PKcs/Artemis Endonuclease Complex Functional Assay

Although the purification described in Chapter 4 was successful in removing most of the contaminations and heterogeneity of the full-length Artemis constructs, dealing with the samples remained challenging. Degradation and aggregation was observed in the different steps during the purification. Therefore, it was important to investigate whether the purified protein was stable, if it was folded properly and biologically active. In the case of wild-type Artemis, the best approach is to carry out a functional assay of the DNA-PKcs/Artemis endonuclease complex as it is the only human endonuclease targeting DNA hairpins.

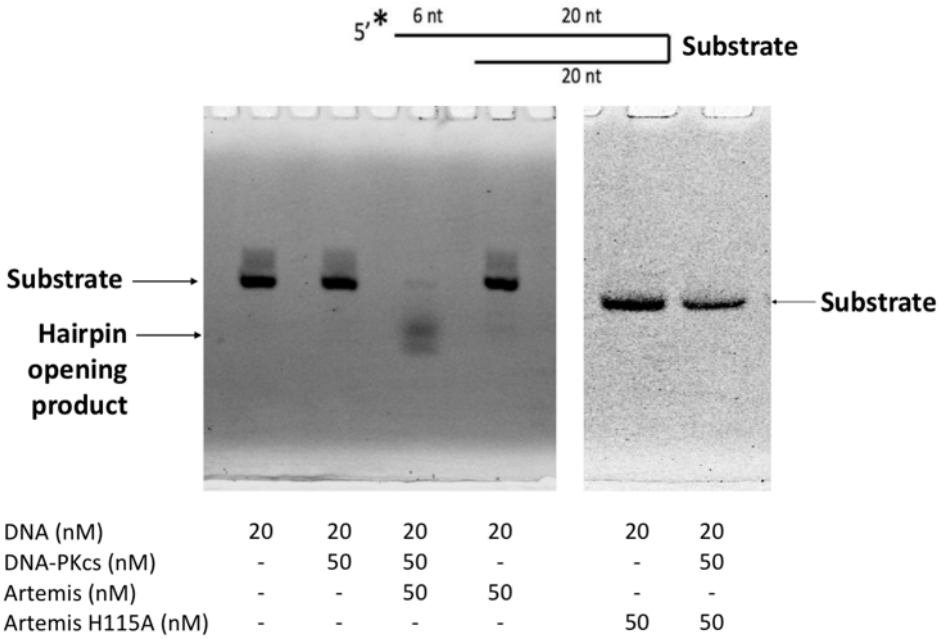


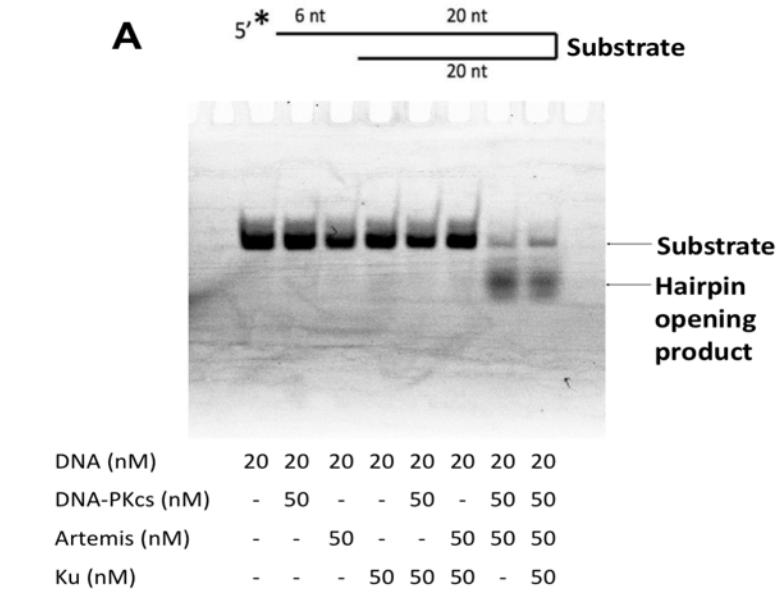
Figure 27. Functional assay of the DNA-PKcs/Artemis endonuclease complexes including the wildtype and mutant targeting the DNA hairpin. The results of the nuclease assay of the DNA-PKcs/wild-type Artemis complex are shown in the gel on the left and those for the DNA-PKcs/ Artemis H115A complex on the right. The substrate is a 20bp dsDNA hairpin with an extra sequence of 6nt ssDNA on the 5' end labelled with cyanine dye 3 (the * label). The gels used for the assay were 12% denaturing polyacrylamide gel. The gels were scanned using Typhoon™ FLA 9000 to detect the fluorescence of the cyanine dye 3 label. The band of substrate and the band of hairpin opening product are indicated and annotated.

For the nuclease assay investigating the hairpin DNA opening activity of DNA-PKcs/Artemis complex, a hairpin DNA labelled with 5' cyanine dye 3 was used as the substrate (Figure 27). When the endonuclease complex is functional, it should cut at the tip of the hairpin DNA to produce a single strand of DNA of around 28nt with the 5' cyanine dye 3 label, which is shorter than the uncut DNA of 46nt.

Figure 27 shows results of the DNA-PKcs/ wild-type Artemis complex cut of the hairpin DNA at the hairpin end where there was a clear signal of the hairpin-opening product and most of the substrate was gone. It was also shown that the endonuclease activity is only active when the wild-type Artemis is in complex with DNA-PKcs. On the other hand, the mutant Artemis H115A is not functional either in complex with DNA-PKcs or not.

Another set of endonuclease-activity assays was then carried out to investigate whether Ku plays a role in modulating DNA-PKcs/Artemis endonuclease complex activity. Ku is known to be the first NHEJ component that initiates the whole process and recruits DNA-PKcs. Besides, there were previous studies indicating interactions among Ku, DNA-PKcs and Artemis. The sample of Ku (full-length Ku70/80 complex) was obtained from Dr Qian Wu and all the possible combinations of Ku, DNA-PKcs and Artemis were tested (Figure 28A). The hairpin DNA was only opened when there was the DNA-PKcs/Artemis complex. Moreover, the presence of Ku slightly reduced the activity of DNA-PKcs/Artemis complex.

Ku has the highest binding affinity to DNA ends. DNA-PKcs is proposed to recruit Artemis to process the DNA ends that require cleaving, including hairpin DNA ends. Therefore, Ku is highly likely to be involved in the reaction when DNA-PKcs/Artemis complex interacts with the hairpin DNA end. However, it is unclear why Ku brings down the endonuclease activity.



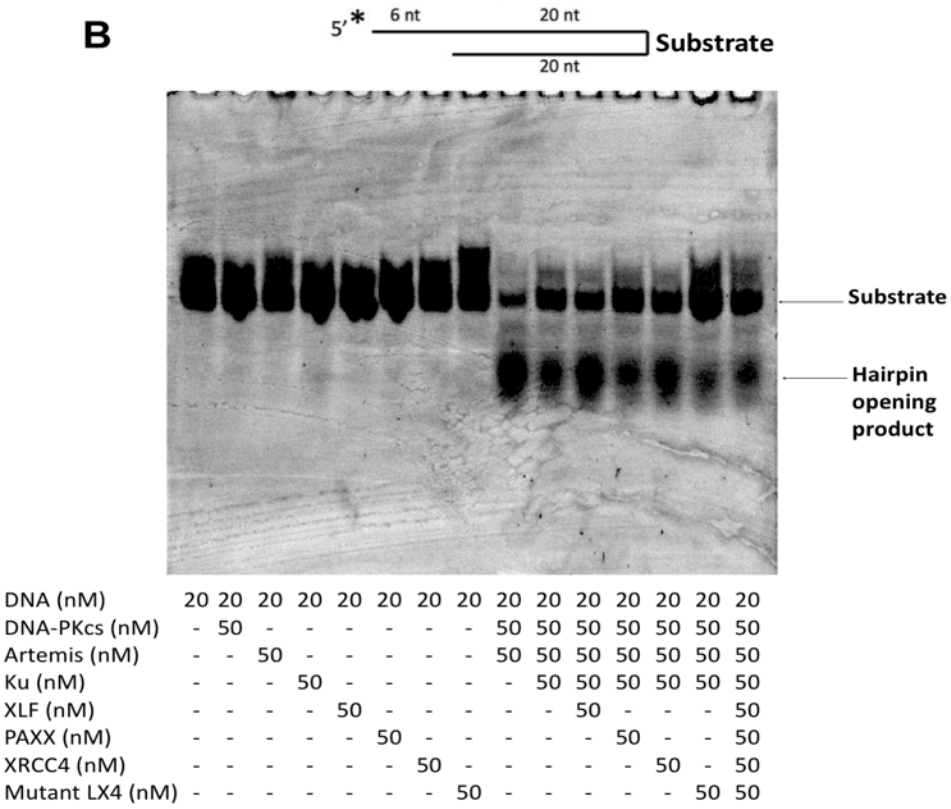


Figure 28. Effect of Ku together with other NHEJ components on the activity of DNA-PKcs/Artemis endonuclease complex targeting the DNA hairpin. (A) Nuclease assay of the DNA-PKcs, wild-type Artemis and Ku. (B) Nuclease assay of the DNA-PKcs, wild-type Artemis, Ku and other NHEJ components. The substrate is a 20bp dsDNA hairpin with an extra sequence of 6nt ssDNA on the 5' end labelled with cyanine dye 3 (the * label). The gels used for the assay were 12% denaturing polyacrylamide gel. The gels were scanned using Typhoon™ FLA 9000 to detect the fluorescence of the cyanine dye 3 label. The band of substrate and the band of hairpin opening product are indicated and annotated.

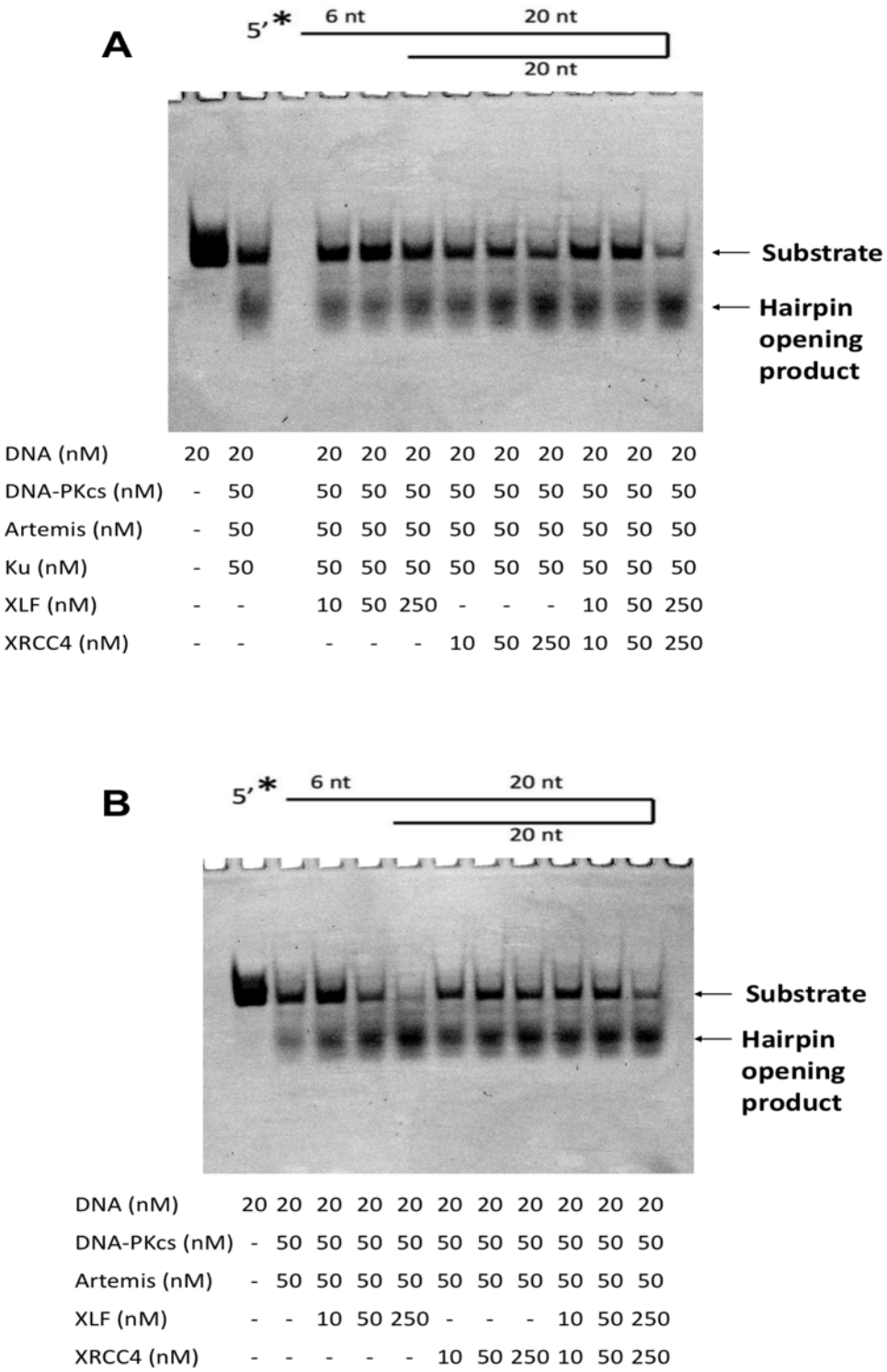


Figure 29. Effect of XLF, XRCC4, XLF/XRCC4 on the activity of DNA-PKcs/Artemis endonuclease complex targeting of the DNA hairpin with and without Ku. (A) Nuclease assay of the DNA-PKcs, wild-type Artemis and Ku. (B) Nuclease assay of the DNA-PKcs, wild-type Artemis, Ku and other NHEJ components. The substrate is a 20bp dsDNA hairpin with an extra sequence of 6nt ssDNA on the 5′ end labelled with cyanine dye 3 (the * label). The gels used for the assay were 12% denaturing polyacrylamide gel. The gels were scanned using Typhoon™ FLA 9000 to detect the fluorescence of the cyanine dye 3 label. The band of substrate and the band of hairpin opening product are indicated and annotated.

There are several hypotheses concerning the reducing effect of Ku. First, it may be that the hairpin DNA is shielded from the endonuclease complex as Ku binds to the DNA strongly. To test this, the nuclease assay with hairpin DNA, Ku, DNA-PKcs and Artemis was done together with a Ku-DNA binding inhibitor from our collaborator (Dr Daruka Mahadevan and Dr Eric Weterings from the University of Arizona). However, the inhibitor did not improve the nuclease activity, suggesting that shielding of the hairpin DNA by Ku should not be the main reason.

Another hypothesis concerning the reducing effect is that other NHEJ components may be needed to improve the efficiency when the system contains Ku, DNA-PKcs and Artemis. To check this hypothesis, the activity of DNA-PKcs/Artemis endonuclease complex was tested with Ku in combination with various NHEJ components that were known to interact with Ku or DNA-PKcs or Artemis including XLF, PAXX, XRCC4 and mutant LX4 [mutant DNA ligase IV complex (LigIV K273A)]. The mutant LX4 instead of the wildtype was used to ensure that the DNA ends would not be ligated to interfere with the results. The protein samples of Ku, XLF and XRCC4 were generous gifts from Dr Qian Wu. According to the results (Figure 28B), Ku continued to interfere with the DNA-PKcs/Artemis endonuclease activity no matter what NHEJ component was added when compared to the group containing only DNA-PKcs/Artemis complex. However, among all the groups containing Ku, DNA-PKcs and Artemis, the one with XLF and the one with XRCC4 clearly had higher efficiency. Interestingly, XLF and XRCC4, belonging to the same XRCC4 superfamily, are known to interact with each other and may form filaments. Therefore, the effects of XLF, XRCC4 and XLF/XRCC4 on the endonuclease activity were further investigated. The proteins were first added to the Ku/DNA-PKcs/Artemis system with a gradient concentration (Figure 29A). It was shown that the increasing concentration of XLF did not significantly increase the endonuclease activity while the increasing amount of XRCC4 stimulated the activity. Moreover, XLF/XRCC4 had the most stimulating effect under the highest concentration.

Furthermore, to test whether the stimulating effect is related to the interaction with Ku, another set of endonuclease assay was done to check the effect of XLF, XRCC4 and XLF/XRCC4 on the DNA-PKcs/Artemis only system (Figure 29B). Interestingly, it was shown that XLF improved the efficiency of the endonuclease complex significantly while XRCC4 was not. The

stimulating effect of XLF/XRCC4 was better than XRCC4 but not comparable to XLF. In conclusion, XLF can significantly improve the efficiency of the DNA-PKcs/Artemis endonuclease complex. However, when Ku is involved, XLF and XRCC4 are both needed to improve the efficiency of the DNA-PKcs/Artemis endonuclease complex.

Another possible reason for the interfering effect of Ku may be that some of the interactions between DNA-PKcs and Ku compete with those between DNA-PKcs and Artemis, reducing the activity of the endonuclease complex. To prove this, more insights from the structural studies are required. In fact, my later cryo-EM structure study of DNA-PKcs/ Artemis complexes structures provided supporting results, which will be introduced in detail in chapter 5.

Last but not least, the interfering effect of Ku can be caused by various factors and the hypotheses mentioned above may all contribute to it. It requires further investigation to understand how Ku is involved in the activity of DNA-PKcs/Artemis endonuclease complex.

5.1.2 Artemis H115A Biophysical Characterisation

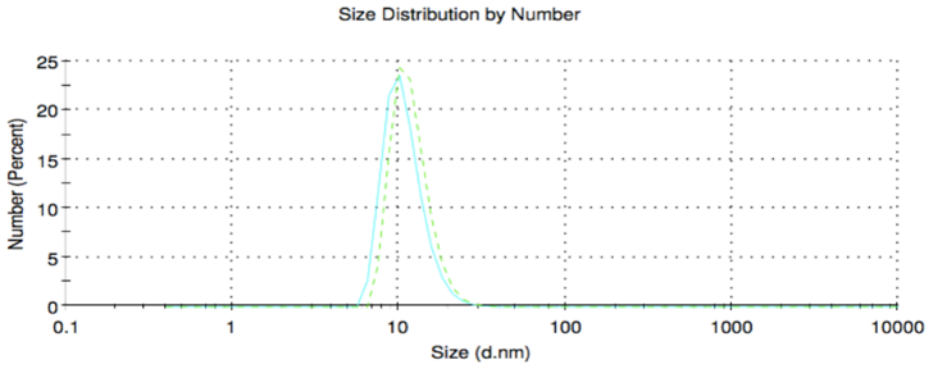
The wild-type Artemis was examined to be physiologic via nuclease assays. However, in the case of Artemis H115A, it is different as the mutant is no longer functional. Therefore, to check the sample heterogeneity and folding, biophysical characterisation using dynamic light scattering (DLS) and circular dichroism (CD) was carried out as a quality check of the sample (Figure 30).

According the DLS results, the sample was rather pure with one peak only (Figure 30A). It suggested that the Artemis H115A sample should be homogeneous in size, which should be around 12nm. This is consistent with the gel-filtration chromatogram (Figure 20C). CD result showed that Artemis H115A was not unstructured (Figure 30B). Moreover, the CD result indicated that the major secondary structure element in Artemis H115A was α-helix. Besides, Artemis H115A was also eluted at a similar position to where the physiologically active wild-type Artemis was eluted in the gel-filtration purification. Therefore, the sample Artemis H115A, which is non-functional, should be properly folded and can be used for structural studies later in complex with DNA-PKcs.

A

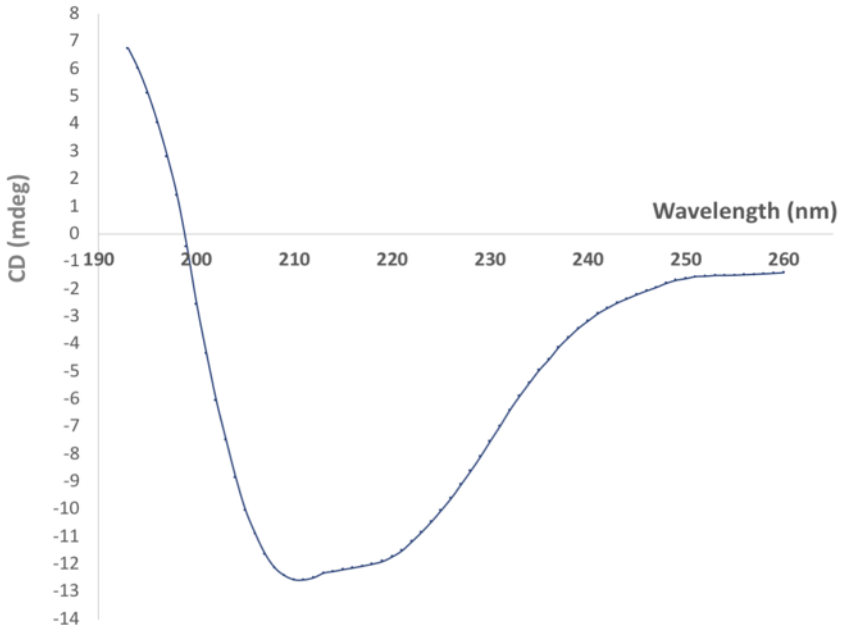
		Size (d.nm):	% Number:	St Dev (d.n...
Z-Average (d.nm): 227.4	Peak 1:	11.97	100.0	3.277
PdI: 0.332	Peak 2:	0.000	0.0	0.000
Intercept: 0.199	Peak 3:	0.000	0.0	0.000

Result quality : Refer to quality report



Dynamic light scattering of Artemis H115A

B



Circular dichroism of Artemis H115A

Figure 30. Biophysical characterisation of Artemis H115A. (A) DLS result of Artemis H115A; (B) CD result of Artemis H115A. The qualitative biophysical characterisation showed that the sample of Artemis H115A should be homogeneous in size, with the diameter of the molecule around 12nm, and folded.

5.1.3 Identification of the Artemis C-terminal Peptide Binding to DNA-PKcs

Based on the sequence alignment, bioinformatics analysis and previous research mentioned in chapter 3, five Artemis C-terminal fragments (Artemis 366-399; 399-408; 385-413; 399-426 and 413-426.) were purified for defining the region interacting with DNA-PKcs. His-tag pulldown assay was carried out on all the fragments with DNA-PKcs.

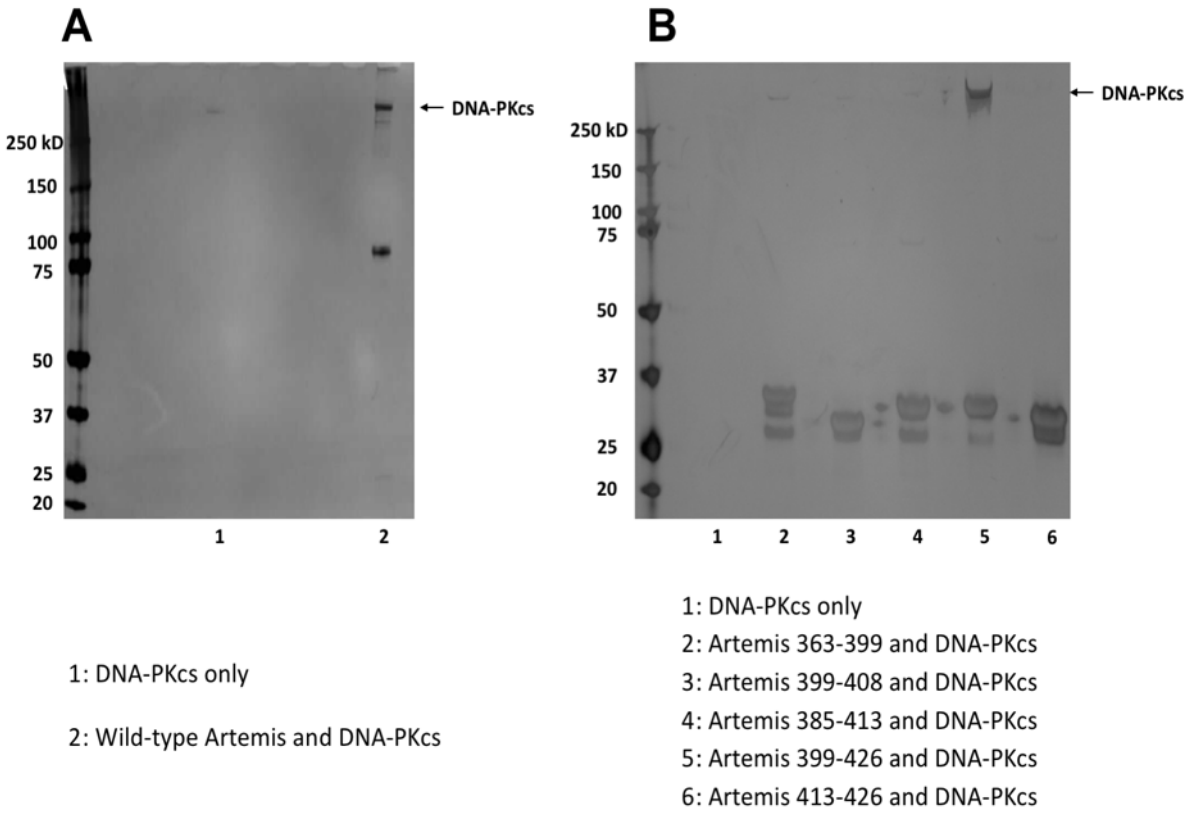


Figure 31. His-tag pulldown assay of DNA-PKcs with different Artemis constructs. (A) His-tag pulldown assay of DNA-PKcs with wild-type Artemis (B) His-tag pulldown assay of DNA-PKcs with Artemis C-terminal peptides including Artemis 363-399; 399-408; 385-413; 399-426; 413-426. The gels were NuPAGE™ 4-12% Bis-Tris Protein Gels stained with silver staining. The numbers shown next to gel are the molecular weight (kDa) of the markers

DNA-PKcs was successfully pulled down with wild-type Artemis as expected, showing that the pulldown experiment is able to confirm the protein-protein interaction (Figure 31A). In the pulldown assays of the Artemis C-terminal peptides, only the fragment containing residues 399-426 is able to pull down DNA-PKcs with a strong signal of the DNA-PKcs band (Figure 31B).

Previously the region of 399-404 was proposed to be the Artemis interaction site with DNA-PKcs. The residues L401 and R402 were specifically highlighted as the double mutant of Artemis L401G+R402N abolished co-immunoprecipitation of endogenous DNA-PKcs (Soubeyrand *et al.,* 2006; Niewolik *et al.,* 2017). However, some of the fragments (Artemis 399-408 and Artemis 385-413) containing the region did not give any positive result. In fact, only Artemis 399-426 was pulled down with DNA-PKcs. It supports the previous bioinformatics analysis in chapter 3 proposing that the region around 420 may also be involved in the DNA-PKcs/ Artemis interaction. It also indicates that, in addition to the interaction from L401 and R402, the interaction from the region between 413-426 is also necessary for the DNA-PKcs/Artemis complex. However, Artemis 413-426 on its own was not able to pull down DNA-PKcs, confirming that the region of 399-404 is also necessary for the interaction.

In conclusion, Artemis 399-426 of the Artemis C-terminal tail is sufficient for the interaction between Artemis and DNA-PKcs. For the first time, the sufficient and necessary region of Artemis interacting with DNA-PKcs is identified. Moreover, this protein-protein interaction is at least a two-patch interaction including the contribution of the region 399-404, which is mainly electrostatic, and the contribution of the region of 413-426, which is likely to be hydrophobic. Subsequent cryo-EM study of the complex of DNA-PKcs/ Artemis 399-426 showed the extra density of the Artemis peptide. It revealed an interesting interaction pattern between DNA-PKcs and Artemis and will be introduced in chapter 6.

5.2 Fragment Based Drug Discovery Initiation on DNA Ligase IV/ Artemis Interaction Site

In addition to the interaction with DNA-PKcs, Artemis is interacting with DNA ligase IV and the structure of the DNA ligase IV/ Artemis complex has been solved (Ochi *et al.,* 2013). The intrinsically disordered peptide of Artemis (485-495) undergoes concerted folding in contact with DNA ligase IV, suggesting that the interaction site should be suitable for drug discovery (Jubb *et al.,* 2015). During my PhD project, the fragment-based drug discovery (FBDD) targeting the DNA ligase IV/ Artemis interaction was initiated (Murray and Blundell, 2010).

FBDD is a promising approach to identify new drugs. Fragments with low molecular weight (usually < 300 Da) can bind hotspots on the target protein-- hotspots are regions within the protein that provide a relatively large contribution towards ligand binding and specific interactions governing the regions (Radoux *et al.*, 2016). Compared to the high-throughput screening (HTS), fewer chemical entities need to be screened and usually more than one fragment binds, allowing choice of starting points for lead discovery. The initial fragment hits in FBDD have lower potency – often in millimolar range - than the more complex molecules in typical HTS compound libraries. However, fragments binding at hotspots with directional and well-defined interactions have high ligand efficiency (Radoux *et al.*, 2016). These fragments are good starting points for linking to other nearby fragments when available or elaboration and fragment growth to any close warmspots, where a fragment may not first bind but where further interactions can stabilize a ligand as it is grown from a fragment bound to a nearby hotspot, to obtain drug-like molecules (Figure 32A) (Thomas *et al.*, 2019).

In the case of DNA ligase IV and Artemis interaction (Figure 32B), there is a potential hotspot with a nearby warmspot where Artemis W489 and F493 interact with DNA ligase IV DBD (Figure 32B), suggesting that this interaction site will be ideal for FBDD. Considering the large consumption on protein sample for initial screening and the stability requirement of the target protein, the DNA ligase IV catalytic domain for which the crystal structure in complex with Artemis peptide was solved, is not perfect for FBDD. Shorter constructs, for example DNA ligase IV DBD, may be better. Structural comparison of the DNA ligase IV catalytic domain

and DBD showed that the structure of the DNA ligase IV/ Artemis interaction site does not change (Figure 32C). Therefore, the DNA ligase IV DBD (DNA ligase IV 230) was used to initiate the FBDD.

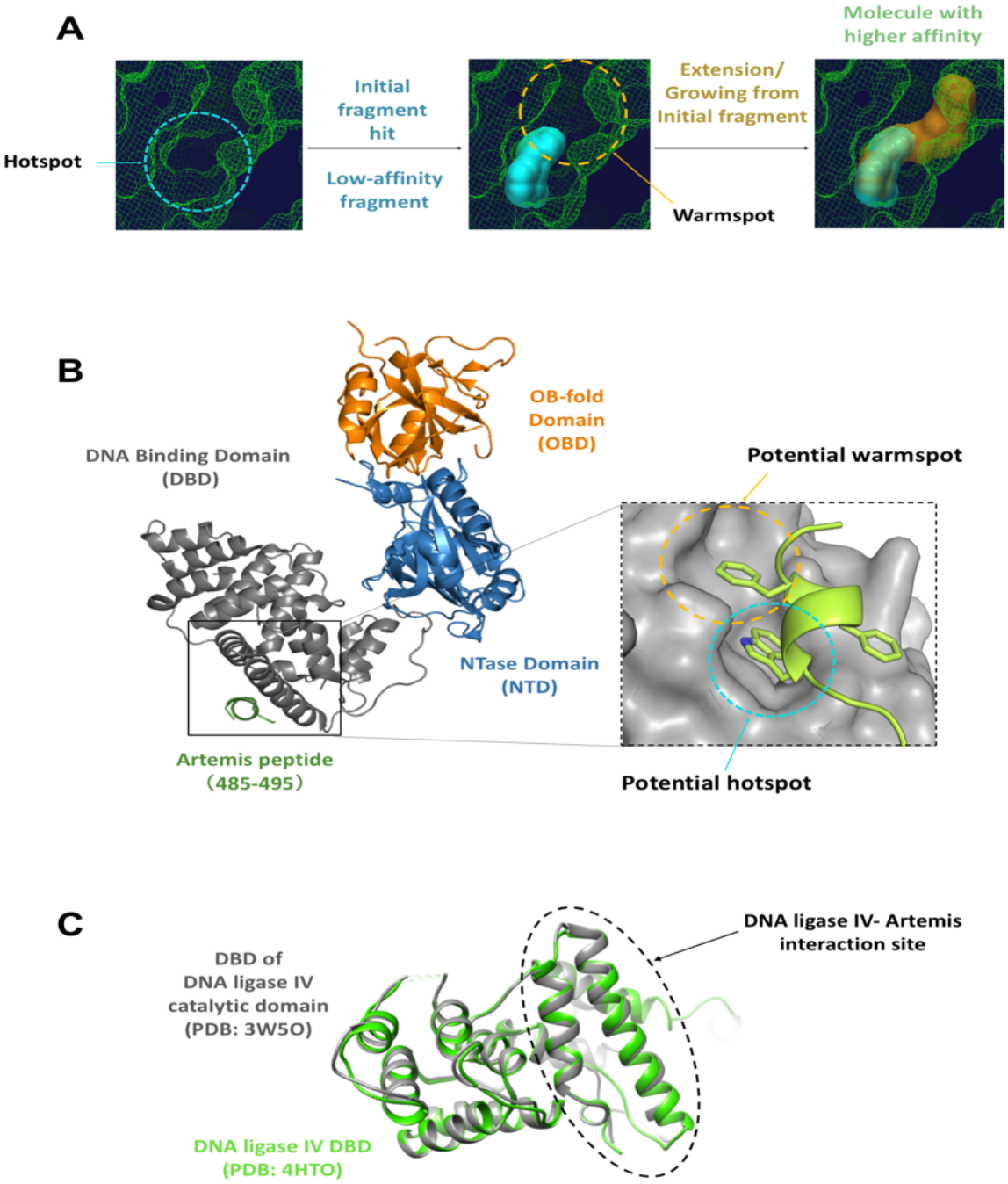


Figure 32. Fragment-based drug discovery (FBDD) for targeting DNA ligase IV/ Artemis interaction site. (A) Schematic diagram of FBDD. The initial fragment hits on the hotspot have high efficiency but low affinity. Subsequent extension from the initial hit to the nearby warmspot can produce more complex drug-like molecule with higher affinity (Figure modified from the presentation of Professor Tom Blundell). (B) Potential FBDD on the interaction site between DNA ligase IV and Artemis, an ideal site for FBDD as there is a hotspot (Artemis W489 docking site) with a nearby warmspot (Artemis F493 docking site) for the initial hit and following fragment extension. (C) Comparison of the Artemis binding site in different structures of different DNA ligase IV constructs. There is no conformational change on the Artemis binding site between the structure of the DBD within the apo DNA ligase IV catalytic domain (grey; PDB: 3W5O) and the structure of apo DNA ligase IV DBD (green; PDB: 4HTO). This suggests that the construct of DNA ligase IV DBD should be suitable for the FBDD targeting the DNA ligase IV/ Artemis interaction site.

As the binding affinities of initial fragments are often low, sensitive screening methods are required. In my case, to have an efficient and sensitive preliminary screening, differential scanning fluorimetry (DSF) or fluorescence-based thermal shift assay was used for the screening of DNA ligase IV 230.

The differential scanning fluorimetry (DSF) or fluorescence-based thermal shift assay is monitoring the thermal unfolding of the target protein with the help of a fluorescent dye that binds to the hydrophobic regions of the unfolded target protein (Figure 33A) (Semisotnov *et al.*, 1991; Pantoliano *et al.*, 2001). It is based on the fact that binding partners/ligands increase the conformational stability of the protein, which is relevant to the change of Gibb's free energy of unfolding (ΔGu). Melting temperature (Tm) of a protein can be defined as the temperature at which folded and unfolded protein are equal with the ΔGu equal to zero. When any ligand from the library enhances the conformation of the target protein, the ΔGu will increase, leading to an increase on the melting temperature (Tm) of the protein.

Fluorescent dyes including Sypro Orange and Nile Red can be used to monitor the thermal unfolding process. Sypro Orange has excitation and emission maxima of 490 nm and 575 nm respectively. It is the most widely used dye for the differential scanning fluorimetry (DSF) or fluorescence-based thermal shift assay and has a high signal-to-noise ratio. The dye is characterised by a low quantum yield (QY)/ low fluorescence when in solvents with high dielectric constants (i.e. the buffer) but obtains high fluorescence/ high QY in solvents with low dielectric constants. Therefore, the low dielectric constant imparted by the melted or unfolded globule state of a protein, where the hydrophobic patches of the protein are exposed, will lead to an increase of the fluorescence signal (Ericsson *et al.*, 2006).

The preliminary differential scanning fluorimetry (DSF) screening of DNA ligase IV 230 involved a library of 960 small molecule fragments. To distinguish the effect of fragments on DNA ligase IV 230, the unfolding curves from separate DSF measurements of the LigIV230 control, without any added fragment, were measured. It indicated a Tm of 53.89±0.36°C of DNA ligase IV 230. Therefore, the analysis of the measured Tm indicates the fragments that induce a positive shift greater than one standard deviation (>0.36°C) or a negative shift lower than one standard deviation (<0.36°C). An example of the analysis of the melting curves with

positive shifts from a screened plate is shown and the full screen with positive hits is available in the supplementary material (Figure 33B, Figure S2).

Multiple fragments had different effects on the melting curves of DNA ligase IV 230 and only the ones inducing a positive shift on the melting temperature were taken as positive hits. In total, there are 24 fragments from the library leading to positive shifts on the melting temperature.

Subsequent steps in FBDD after initial fragment screening include fragment hit validation and characterisation. To understand the mode of action of the fragment, crystal structure with the fragment binding is important and it is crucial to obtain apo DNA ligase IV 230 crystal for fragment soaking. Currently the crystallisation is under optimisation.

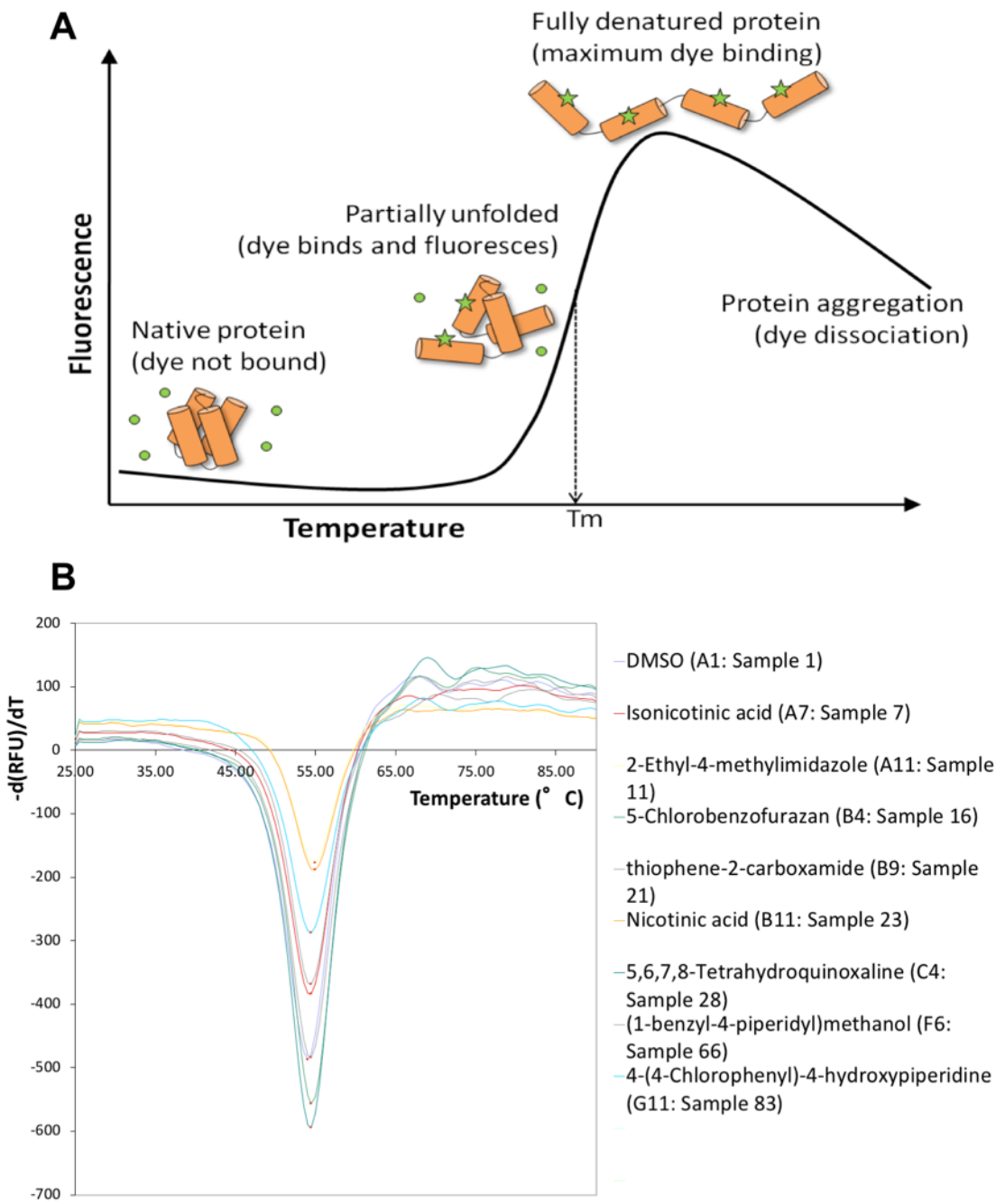


Figure 33. Thermal Shift Assay/ DSF of fragment screening and an example of DNA ligase IV DBD. (A) Schematic diagram of the basic principle of fluorescence-based thermal shift assay. The native protein (orange) does not bind to the dye (green) because of the high dielectric constant of the solvent. As the temperature increases, the protein becomes more unfolded, exposing the hydrophobic regions to bind the dye and resulting in increasing fluorescence signal. Melting temperature (Tm) of a protein can be defined as the equilibrium temperature, under which the folded and unfolded protein are in equal amounts. (B) An example Thermal Shift Assay/ DSF profile of fragment screening on DNA ligase IV DBD. The example is the plate 7 of the NMR fragment library. Only the positive shifts in melting temperature on binding with fragments are shown in the example.

5.3 Temporal Organisation of NHEJ

The spatial organisation of NHEJ can be studied via biochemical and structural studies of NHEJ complexes. However, in addition to the spatial organisation of NHEJ, the temporal organisation has also been of interest to the group and always been a key question in the field. Although the three main steps of DNA end recognition, end synapsis and processing and end ligation were known, it has been unclear at which points of the process different NHEJ components participate. The full timeline remains amorphous especially within the step of end synapsis and processing. Structural studies can help us understand the timeline but it is relatively challenging. In collaboration with the Strick group, we applied the technique of single-molecule study to the NHEJ system. Moreover, using a novel DNA molecular forceps, we were able to visualise and understand the temporal organisation of NHEJ for the first time.

To study quantitatively the properties of complex molecular synapses and interactions, a DNA-scaffold-based single-molecule assay was developed to observe repeated cycles of synapsis and rupture. Microscopic properties including lifetime of synapsis and kinetics of formation can be derived from such assays (Figure 34A). The DNA molecular forceps is made up of two linear dsDNA segments around 1510 bp in length, connected to each other by a third double strand segment, termed “bridge,” of around 690 bp. The bridge is docked at the point that is 58 bp away from the ends of the DNA segments it tethers, allowing the ends to fluctuate freely and providing enough space of DNA for the loading of NHEJ components. One end of the DNA construct (around 3.6 kbp) is attached to a glass surface and the other end to a magnetic bead. This system enables real-time recording of DNA end synapsis/ligation by monitoring the molecule’s overall extension. If there is no synapsis/ligation between the ends, the extension of the 3.6 kbp construct will be around 1085 nm with the two linear dsDNA segments and bridge pulled to their full length. On the other hand, if there is any synapsis/ligation between the ends, the linking bridge will not be involved in the extension of the DNA forceps and the extension of the 3.6 kbp construct will be around 913 nm, which is 172 nm shorter than 1085 nm of the extension with no synapsis/ligation (Figure 34B).

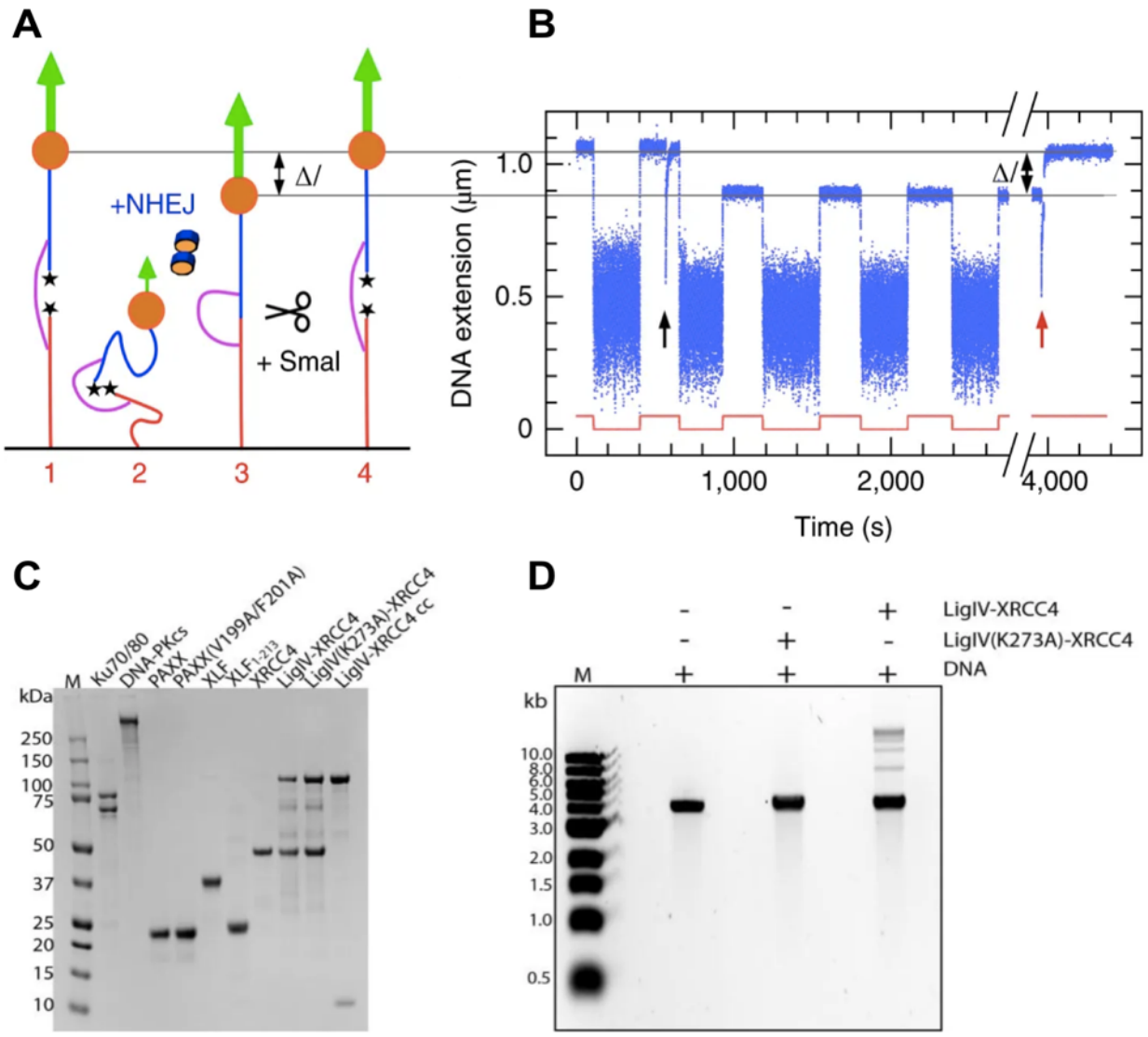


Figure 34. Experimental design of studying NEHJ temporal organisation with the DNA forceps. (A) Schematic diagram of the single-molecule method using the DNA forceps with full NHEJ components leading to DNA ligation. Two 1.5 kbp dsDNA segments (blue, red) are joined by a 600 bp dsDNA "bridge" segment (magenta) to form the DNA construct used in this study. Stars at the DNA ends represent phosphate groups. One end of the construct is tethered to the bottom glass surface and the other end is tethered to a 1-micron magnetic bead. A pair of magnets located above the beads allows one to generate a controlled, vertical extending force on the DNA (green arrow). The resulting end-to-end extension of the DNA is determined by computer-based monitoring of the bead position above the surface. DNA ligation is observed following four steps: (1) at high force and in the absence of end-end interactions, a high-extension state is observed, (2) the force is lowered to allow the ends to encounter and the interaction of NHEJ to process, (3) the force is increase but if there has been end ligation, the construct cannot return to the full extension, (4) The full extension is recovered upon cleavage of SmaI. This system can be used repeatedly; (B) Real-time monitoring of the DNA extension upon the modulation of force (bottom red line) in the presence of full-length Ku70/80, DNA-PKcs, PAXX, XRCC4, XLF, and DNA ligase IV complex. At the beginning, DNA extension (blue points) is shown to switch between a low and a high value when force is modulated. Upon addition of NHEJ components (black up arrow), the extension under high-force situation is later reduced (Δl) due to the end ligation. After the addition of SmaI restriction enzyme (red up arrow), there is a sudden increase of the DNA extension (Δl) and the DNA extension is back to the initial full length;

(C) Samples of the NHEJ components involved in the single-molecule study on a NuPAGE™ 4-12% Bis-Tris Protein Gels stained with Coomassie blue. The numbers shown next to gel are the molecular weights (kDa) of the markers. Among all the components, I purified and provided the samples of DNA-PKcs, PAXX constructs and DNA ligase IV constructs. (D) DNA ligation assays with DNA ligase IV complex and its catalytically dead mutant on a 1% agarose gel stained with SYBR gold. The DNA ligase IV complex is functional and the mutant complex is catalytically dead. The numbers shown next to gel is the ladder with DNAs of different lengths (kb).

To test and validate the system, full-range NHEJ components were first applied including full-length Ku, DNA-PKcs, PAXX, XRCC4, XLF, and DNA ligase IV complex. All the samples used for the single-molecule experiment are pure and the DNA ligase IV complex is physiologic (Figure 34C D). The system was able to recapitulate the complete NHEJ as the ligation took place and a stable and lasting end ligation was monitored from the length of DNA extension under high magnetic force (Figure 34B)-- After adding NHEJ component and the incubation at low force, there was a decrease of the extension length (ΔI), which was around 170 nm, suggesting that the ends should be ligated. Moreover, this could be inverted by adding the restriction enzyme of SmaI to cleave the ligated DNA.

Stepwise assembly was then examined with the system to understand the temporal organisation of NHEJ especially during the step of DNA end synapsis. Ku and DNA-PKcs should always be there for DNA end recognition. It was found that the combination of the DNA-PK holoenzyme (Ku70/80 and DNA-PKcs) and PAXX is the minimal requirement to observe consistent and frequent yet short-lived interactions between the two DNA ends (Figure 35A). End synapsis was observed as a transient plateau in the DNA extension signal obtained after switching the force from F_{low} to F_{high}; The lifetime of the end synapsis, reflected by the duration of the plateau, was around 2.2 s and the change of the extension of the DNA upon rupture was around 166 nm (Figure 35A). However, this DNA synapsis interaction disappeared when the PAXX mutant (PAXX V199A F201A) was used instead of the wildtype (Figure 35B). The PAXX mutant is not able to interact with Ku, indicating that the interaction with Ku is vital for the synapsis. Nevertheless, adding XLF, which also interacts with Ku and belongs to the same XRCC4 superfamily with PAXX, did not form any end synapsis. Moreover, neither the combination of Ku and DNA-PKcs nor PAXX on its own could form the previous synapsis

(Figure 35C). This suggests that Ku, DNA-PKcs and PAXX should be the "upstream" components of NHEJ that enable the primary formation of DNA end synapsis.

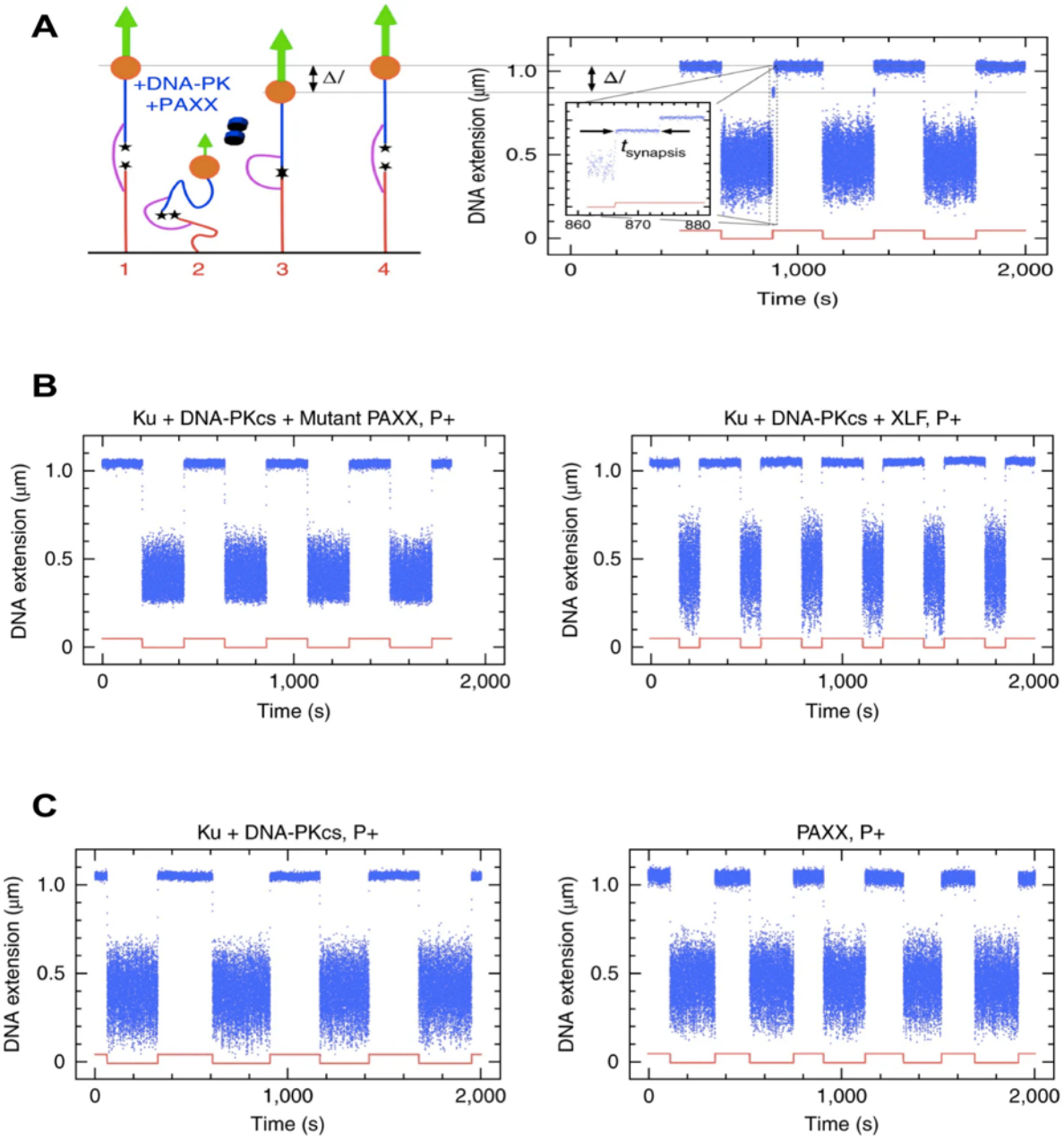


Figure 35. Ku, DNA-PKcs and PAXX are the minimal combination to form consistent and stable yet short-lived DNA end synapsis. (A) Schematic diagram of the experimental design for monitoring end synapsis with Ku, DNA-PKcs and PAXX and the real-time monitoring of the DNA extension. Stars at the DNA ends represent phosphate groups. There is a clear plateau representing the end synapsis when switching from low force to high force; (B) Real-time monitoring of the DNA extension of the reaction system of Ku, DNA-PKcs, and the PAXX mutant (left panel) and the system of Ku, DNA-PKcs, and XLF (right panel); (C) Real-time monitoring of the DNA extension of the reaction system of Ku and DNA-PKcs (left panel) and the system of PAXX only (right panel). The DNA end synapsis is only happening in the system of Ku, DNA-PKcs, and PAXX, indicating the Ku, DNA-PKcs and PAXX are the upstream NHEJ components for end synapsis.

Other NHEJ components including XLF, XRCC4 and DNA ligase IV complex were then added to the "upstream" NHEJ system in different combinations. In general, the full NHEJ system with Ku, DNA-PKcs, PAXX, XLF, XRCC4 and DNA ligase IV complex has the longest DNA synapsis lifetime of around 66 s (Figure 36A). In this case, the DNA ends were dephosphorylated to prevent DNA end ligation. Getting rid of DNA ligase IV complex or XLF or XRCC4 reduced the lifetime of synapsis to around 2 s, which should be the primary synapsis formed by Ku, DNA-PKcs and PAXX. This further confirmed the upstream effect of Ku, DNA-PKcs and PAXX and indicated that XLF, XRCC4 and DNA ligase IV complex are likely to be involved as one functional unit in the end-synapsis step. Therefore, the unit of XLF, XRCC4 and DNA ligase IV complex was also added to the system of Ku and DNA-PKcs to test the function. Interestingly, end synapsis was observed with a lifetime of 9 s. It should be noted that Ku and DNA-PKcs must be included in the system or there will be no synapsis at all. Therefore, the functional unit of XLF, XRCC4 and DNA ligase IV complex must act on the base of DNA-PK.

In general, based on all the combinations we had, a model of the NEHJ end synapsis was proposed (Figure 36B). Ku and DNA-PKcs first arrive at the DNA ends in the end-recognition step. The recruitment of PAXX and its interaction with Ku helps form short-lived but consistent DNA end synapsis, which is likely to be the upstream reaction in the end-synapsis step. Meanwhile, the functional unit of XLF, XRCC4 and DNA ligase IV complex can also help form DNA end synapsis together with DNA-PK with a longer lifetime. Together, in the full system of all NHEJ components, the robust and long-lived end synapsis can be formed.

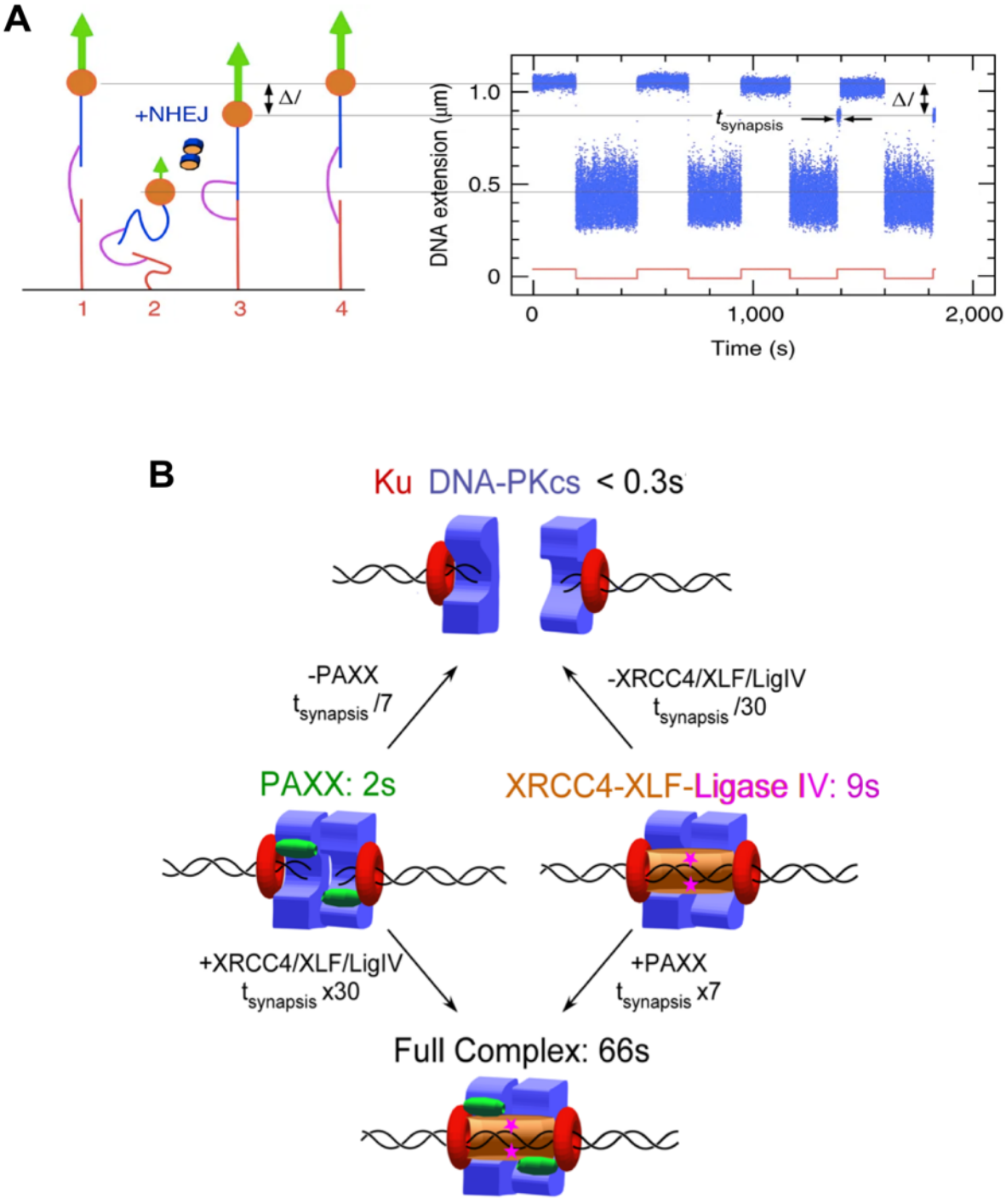


Figure 36. The effect of the full NHEJ complex in end synapsis and hypothesis of the NHEJ end synapsis process. (A) Schematic diagram of the experiment design of monitoring end synapsis with Ku, DNA-PKcs, PAXX, XLF, XRCC4 and DNA ligase IV complex and the real-time monitoring of the DNA extension. The DNA ends are dephosphorylated to prevent end ligation. There is a longer plateau representing the end synapsis compared to that in the system of Ku, DNA-PKcs and PAXX; (B) Model of the organisation of NHEJ end synapsis. The initial DNA-PK holoenzyme complex with a putative lifetime in the range of hundreds of milliseconds, which can be stabilized by the incorporation of PAXX (left) or XRCC4/XLF/DNA ligase IV complex (right). A complete complex stabilized by both PAXX and XRCC4/XLF/Ligase IV has the longest lifetime (bottom).

5.4 Summary

The biochemical and biophysical characterisations facilitated findings that had not been previously observed.

First, the assays of DNA-PKcs/ Artemis endonuclease complex, showing that the purified sample is functional, demonstrated that other NHEJ components Ku can inhibit the nuclease activity while XRCC4 and XLF/XRCC4 complex can reduce the level of inhibition in a concentration-dependent manner. Moreover, XLF strongly stimulates the endonuclease activity of DNA-PKcs/Artemis complex.

Secondly, although the Artemis H115A is not active, it is structured and should be good for further structural studies.

Moreover, the pulldown experiment identified the region of Artemis interacting with DNA-PKcs for the first time. More importantly, it provides a stage for the subsequent cryo-EM structural study of DNA-PKcs/ Artemis complex.

Besides, initial fragment screening showed that FBDD was practical targeting the interaction site of DNA ligase IV and Artemis with 24 fragments showing positive results in the initial DSF screening.

Last but not least, the single-molecule methods revealed the temporal organisation of NHEJ especially in the step of end synapsis. PAXX was first shown to have a specific role in NHEJ to form very early end synapsis. In addition, XRCC4/XLF/DNA ligase IV was shown to be one function unit in end synapsis. Together, they could form strong and long-lasting end synapsis for further NHEJ process.

Chapter 6. Cryo-EM of DNA-PKcs/Artemis Related Complexes

In Chapters 4 and 5 I have discussed the sequence of Artemis interacting with DNA-PKcs and the effects of other NHEJ components on the endonuclease activity of DNA-PKcs/ Artemis complex. However, to understand how DNA-PKcs and Artemis interact in detail, a structural study is necessary. Considering the flexibility and complexity of the complex and encouraged by the resolution revolution that has developed since 2013 in cryo-EM, I decided using cryo-EM as the main structural method to study the DNA-PKcs/ Artemis complex.

In fact, before the resolution revolution, there had already been many EM studies targeting the DNA-PKcs and its complexes. The initial studies can be traced back more than 20 years when EM and atomic force microscopy (AFM) were first used to study the structure of DNA-PK (Cary *et al.,* 1997; Yaneva *et al.,* 1997). Later cryo-EM imaging of the catalytic subunit of the DNA-PKcs was successful in providing an image at resolution of 22 Å (Chiu *et al.,* 1998), followed by another map at a similar resolution using electron crystallography (K. Leuther *et al.,* 1999). Further EM studies on DNA-PKcs-related complexes were reported at the turn of the century, when it was shown that kinase activity of DNA-PKcs is related to DNA concentration and EM work revealed that two DNA-PKcs molecules can form complexes, which may play a role in DNA end synapsis (DeFazio *et al.,* 2002). Later work from Spagnolo et al. (2006) further reinforced this point, presenting the structure of the DNA-PK assembled on DNA and discussed its implications for NHEJ. In addition, there were many EM studies at that time targeting DNA-PKcs and related complexes but the highest resolution was 7 Å (Boskovic *et al.*, 2003; Rivera-Calzada *et al.*, 2005, 2007; Williams *et al.*, 2008). Since the resolution revolution, more cryo-EM studies have been conducted and the resolution of DNA-PKcs was pushed to 4.4 Å, DNA-PK to 6.6 Å (Sharif *et al.,* 2017; Yin *et al.,* 2017).

During the past four years, I have conducted cryo-EM studies of different constructs of DNA-PKcs/ Artemis related complexes. Before starting the cryo-EM work, negative staining was conducted to check the sample quality of DNA-PKcs and full-length Artemis. Later, the first

cryo-EM experiments targeted DNA-PKcs in complex with wild-type Artemis, followed by the study of the protein/ DNA complex including DNA-PKcs, Artemis H115A and hpDNA. Last but not least, the structure of DNA-PKcs/ Artemis C-terminal peptide complex was studied and provided the highest resolution of 4.2 Å, clearly showing the extra density of Artemis peptide and revealing the interaction pattern between DNA-PKcs and Artemis. In parallel I have carried out a study of apo-DNA-PKcs, to provide the best structure of the molecule for further complex study and have achieved an improved map at 3.88 Å resolution, somewhat higher resolution than the earlier X-ray studies in the Blundell lab (Sibanda *et al.,* 2017).

6.1 Negative Staining of DNA-PKcs and Artemis

Before initiating cryo-EM experiments, preliminary screening using negative staining was conducted to check the sample quality of DNA-PKcs and Artemis, in addition to the previous biophysical characterisation. Compared to cryo-EM, there are several advantages of negative staining. First, negative staining requires much less protein - ten-to-a-hundred-fold less - compared to cryo-EM, thanks to the supporting film on the grid. Second, negative staining

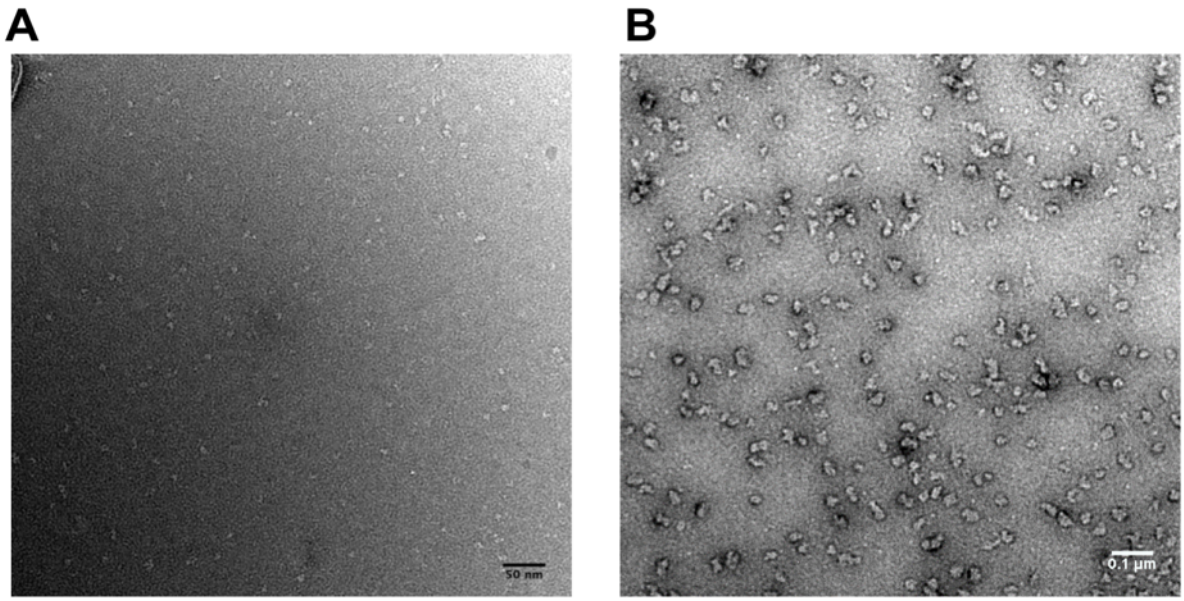


Figure 37. Negative-staining screening of Artemis and DNA-PKcs. (A) Negative staining of wild-type Artemis. The protein appeared to be homogeneous in size with various different shapes, indicating the different 2D projections and probably also flexibility of the sample; (B) Negative staining of DNA-PKcs. The protein particles are similar in size. Most of the particles were single while come of the molecules were adhering to others.

provides stronger contrast than cryo-EM. This is very helpful for visualising protein with low molecular weight, which is exactly the case of Artemis. Third, the sample preparation is much easier and faster, with no facility requirement, and the grids can be stored at room temperature for months. Fourth, the microscope requirement is lower and costs less.

In the negative-staining micrograph of wild-type Artemis (Figure 37A), there are many single particles of Artemis, of the same size, evenly distributed. This suggests that the purified sample of Artemis should be homogeneous in the storage buffer. Moreover, it is noticed that, although the particles are similar in size, the shapes of the particles are relatively different. This may be due to the various views of the protein or the result of the flexibility of Artemis.

In the negative-staining micrograph of DNA-PKcs (Figure 37B), the particles of DNA-PKcs are also evenly distributed, proving the homogeneity of the purified sample. Unlike Artemis, DNA-PKcs appears to have a relatively fixed structure, as observed in previous studies. Moreover, although there are many nicely separated single particles, some of the particles also interacted with others, forming complexes with two or more molecules involved.

In summary, the negative-staining results showed that the purified protein samples were nicely distributed without aggregation and should be good for subsequent cryo-EM experiments, which are crucial for obtaining high-resolution information.

6.2 Cryo-EM of DNA-PKcs/Artemis Complex

A cryo-EM investigation was initiated targeting the structure of the DNA-PKcs/ Artemis complex, a protein-protein complex, to see how the full-length proteins interact with each other and whether there are any previously undiscovered interactions in addition to that of DNA-PKcs with Artemis C-terminal peptide (399-426).

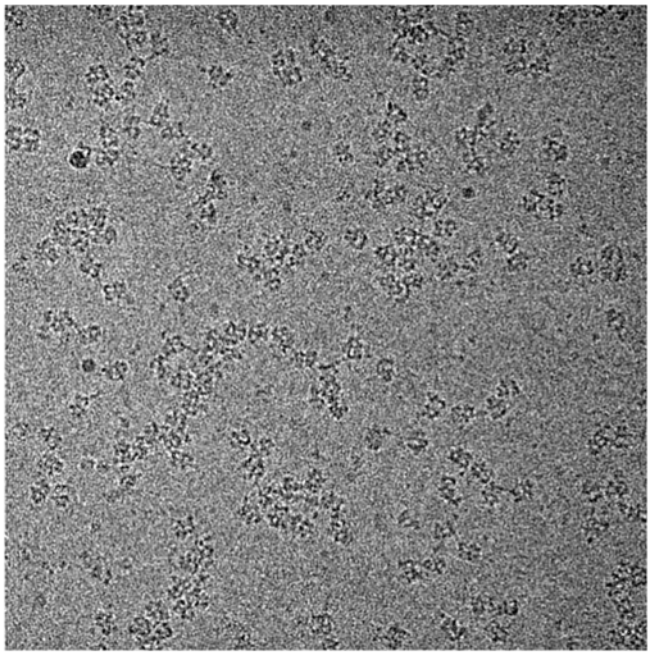
6.2.1 Sample preparation, grid screening and data collection

The purified DNA-PKcs, at a final concentration of 0.3 mg/ml, was added to wild-type Artemis with a 1:2 ratio. The final buffer was the same as the nuclease assay buffer. Vitrobot was used for the grid preparation and aliquots of 2.5 µl sample were applied to glow-discharged grids (Quantifoil Cu 1.2/1.3, 300 mesh), which were then blotted for 2.5 s with filter paper and plunge-frozen in liquid ethane.

The grids were first screened and an example of the micrographs is shown in Figure 38. According to the micrographs, there are a few interesting features regarding the behaviour of the sample. First, some particles remain single with no attachment to other molecules while others are inter-connected and form long curved filaments. Moreover, there are various visible strings linked with the particles of DNA-PKcs, which are very likely to be Artemis protein. Some of the strings behave like a tail of DNA-PKcs, floating in the buffer, while others connect different particles of DNA-PKcs. This indicates that Artemis may have the potential to link DNA-PKcs together, although it is unclear whether the linkages and filaments are due to interactions between DNA-PKcs and Artemis or DNA-PKcs and DNA-PKcs or Artemis and Artemis or all three.

To obtain more information at higher resolution, the grid was viewed on a Thermo Fisher Titan Krios electron microscope (300kV) with a Gatan K2 detector and the movies were recorded under the counting mode with a magnification of 130,000 times. The pixel size used for the collection was 1.07 Å. Total dose for a movie stack was 49.4 electrons/$Å^2$ with an exposure time of 14 s. A series of defocus value were used for collection including -1.3, -1.6,

-1.9, -2.2, -2.5, -2.8, and -3.1 µm. 3303 movies were collected at the end of the data collection and processed subsequently.



Microscope	Titan Krios
Detector	K2 (counting)
Nominal Magnification	130,000x
Pixel Size, Å per pixel	1.07
Total Dose, electrons/Å^2	49.4
Exposure, sec	14
Defocus Range, µm	-1.3 ;-1.6; -1.9; -2.2; -2.5; -2.8; -3.1

Figure 38. Cryo-EM screening and data collection of DNA-PKcs/Artemis complex. The left panel is an example of a micrograph of the grids of DNA-PKcs/Artemis complex. There are evenly distributed particles in the micrograph. The particles have the expected shape of DNA-PKcs, as it is the largest component in the complex, around six-time the molecular weight of Artemis. There are also many strings attached to some of the particles, which are very likely to be Artemis. Some of the particles are single with or without a tail-like string floating around, while others are inter-connected. The right panel shows the parameters that were used for the final data collection. The dataset was collected at the cryo-EM facility of the Department of Biochemistry, University of Cambridge.

6.2.2 Data processing

Cryo-EM data processing is a rapidly developing field. As mentioned in section 1.4.3, there are many software packages for single particle analysis and Relion has been the mostly used one during my PhD.

The classical data processing procedure includes: motion correction, CTF estimation, particle picking, 2D classification, 3D classification, 3D refinement and post-processing (Scheres, 2016). In general, the objective of the processing is to achieve three ultimate goals: picking clean and nice particles without contamination, accurate classification of particles of similar conformations and precise determination of projection angles for 3D reconstruction from 2D images. Therefore, the pathway for obtaining the EM map is not restricted to the classic one and can be altered to achieve a better map.

Various pathways of data processing for DNA-PKcs/Artemis complex were attempted. The best map achieved was at the resolution of 6.2 Å following the pathway shown in Figure 39. The 3303 collected micrographs were first sent for motion correction and CTF estimation. 3140 micrographs with good CTFs were picked and used for particles auto-picking. 110759 particles were picked and extracted from the micrographs and sent for 3D classification directly into 8 classes. 3 classes accounted for most of the particles including classes 5, 6 and 7. When compared to the cryo-EM map of DNA-PKcs, only class 7 showed clear extra densities. Class 7 was then subjected to another 3D classification with class 1 having 96.2% of the particles, filtering out the contamination and bad particles in class 2 (3.8%). Particles of class 1 then went through 3D refinement and postprocessing to obtain the final map of 6.2 Å.

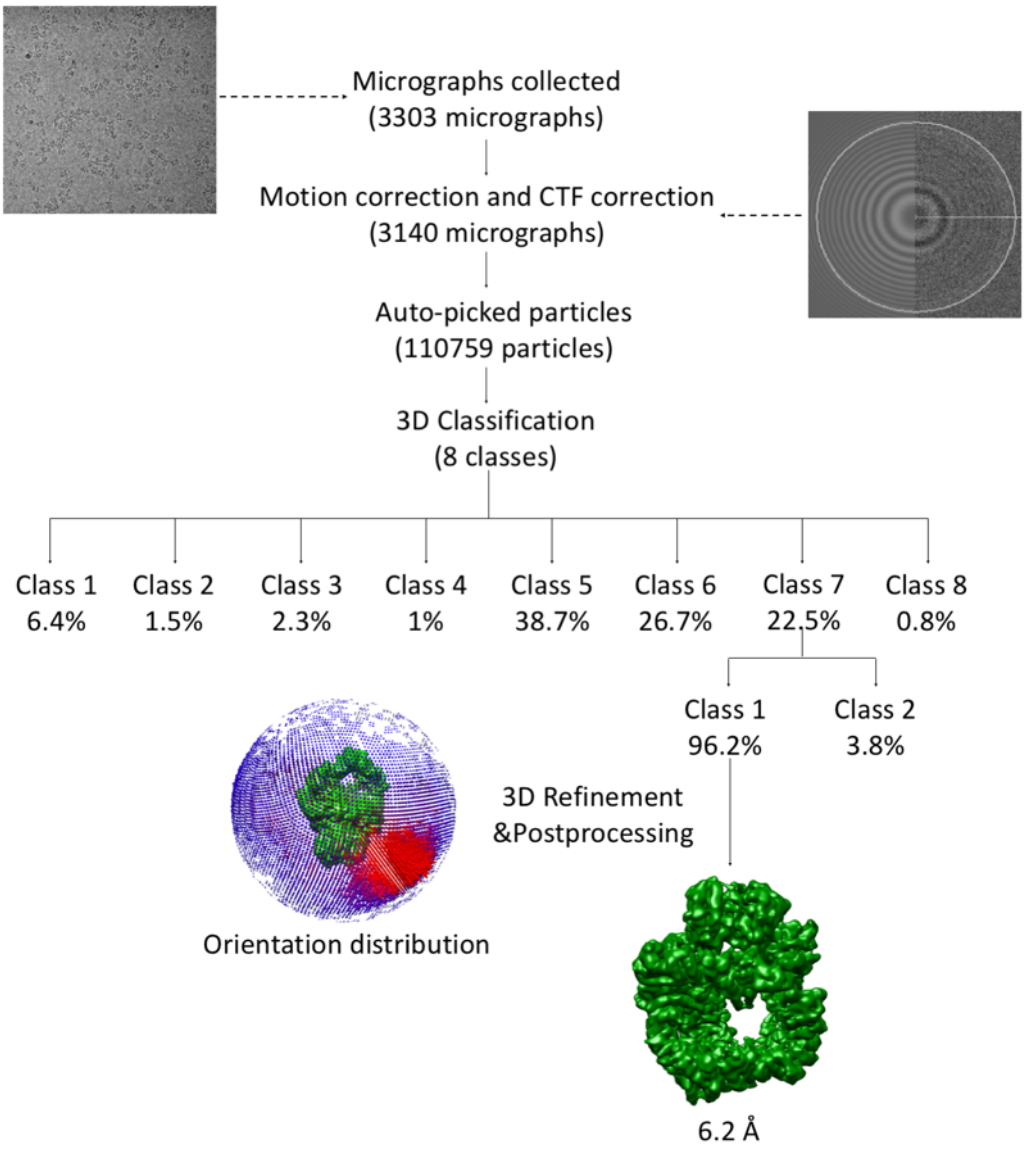


Figure 39. Cryo-EM data processing pathway of the map of DNA-PKcs/ Artemis complex at the resolution of 6.2 Å. Unlike the classic data process, 2D classification was not involved for the reconstruction of the map. 3140 micrographs were picked from the original 3303 collected after motion correction and CTF estimation. 110759 particles were picked and extracted. 3D classification for 8 classes of those particles cleaned up most of the contamination from the nice particles and class 7 contained extra density compared to the apo DNA-PKcs map. Class 7 was then further classified in to two classes allowing filtering out 3.8% of the particles. The final class went through 3D refinement and postprocessing to achieve a cryo-EM map of DNA-PKcs/ Artemis complex at the resolution of 6.2 Å.

6.2.3 Cryo-EM map analysis

Generally, the cryo-EM map of DNA-PKcs/ Artemis complex has a similar overall shape to DNA-PKcs (Figure 40A). Of particular note, the N-terminal arm looks very different. It was known that the N-terminal arm is flexible and has continuous up-and-down movements. Therefore, the density of the N-terminal arm was relatively weak and hard to describe in both the cryo-EM and the crystal structure (Sharif *et al.,* 2017; Sibanda *et al.,* 2010; Sibanda *et al.,* 2017). In the map of the DNA-PKcs/ Artemis complex, the N-terminal arm region has more density than the apo-DNA-PKcs (Figure 40B site B). Moreover, the extra density builds up on the original N-terminal arm, as it approaches the FAT domain. Although the somewhat fragmented density is unlikely to allow association with the sequence, it is clear that the extra density in the N-terminal region results from DNA-PKcs/ Artemis interactions. Considering the flexibility of the density, it should come from the C-terminal tail of Artemis instead of the structured nuclease region.

There are also two further differences between DNA-PKcs/ Artemis complex and DNA-PKcs (Figure 40B site A site C). Site A sits close to the polypeptide chain between the circular cradle and kinase head region. As mentioned in section 1.3.2.2, this is around residue 2576-2744 of DNA-PKcs, which covers the ABCDE cluster and has density missing. In the density map of X-ray crystallisation, part of the region is observed and four-α-helices (probably representing 2602-2665) are placed in the model hanging down into the central cavity of DNA-PKcs molecule. In all the available cryo-EM maps of DNA-PKcs (Sharif *et al.*, 2017; Yin *et al.*, 2017; Wu *et al.*, 2019), there is no density observed in this region. However, in the map of DNA-PKcs/ Artemis complex, there are extra densities in site A and they are very likely to be part of the missing 2576-2744 due to the continuity of the EM map. This suggests that the interaction between Artemis and DNA-PKcs may have an impact on the conformation of DNA-PKcs and stabilise the flexible region covering ABCDE cluster. In addition, there is extra density in DNA-PKcs/ Artemis complex at site C compared to apo DNA-PKcs, which is at a similar position where Ku binds (Sharif *et al.*, 2017; Yin *et al.*, 2017).

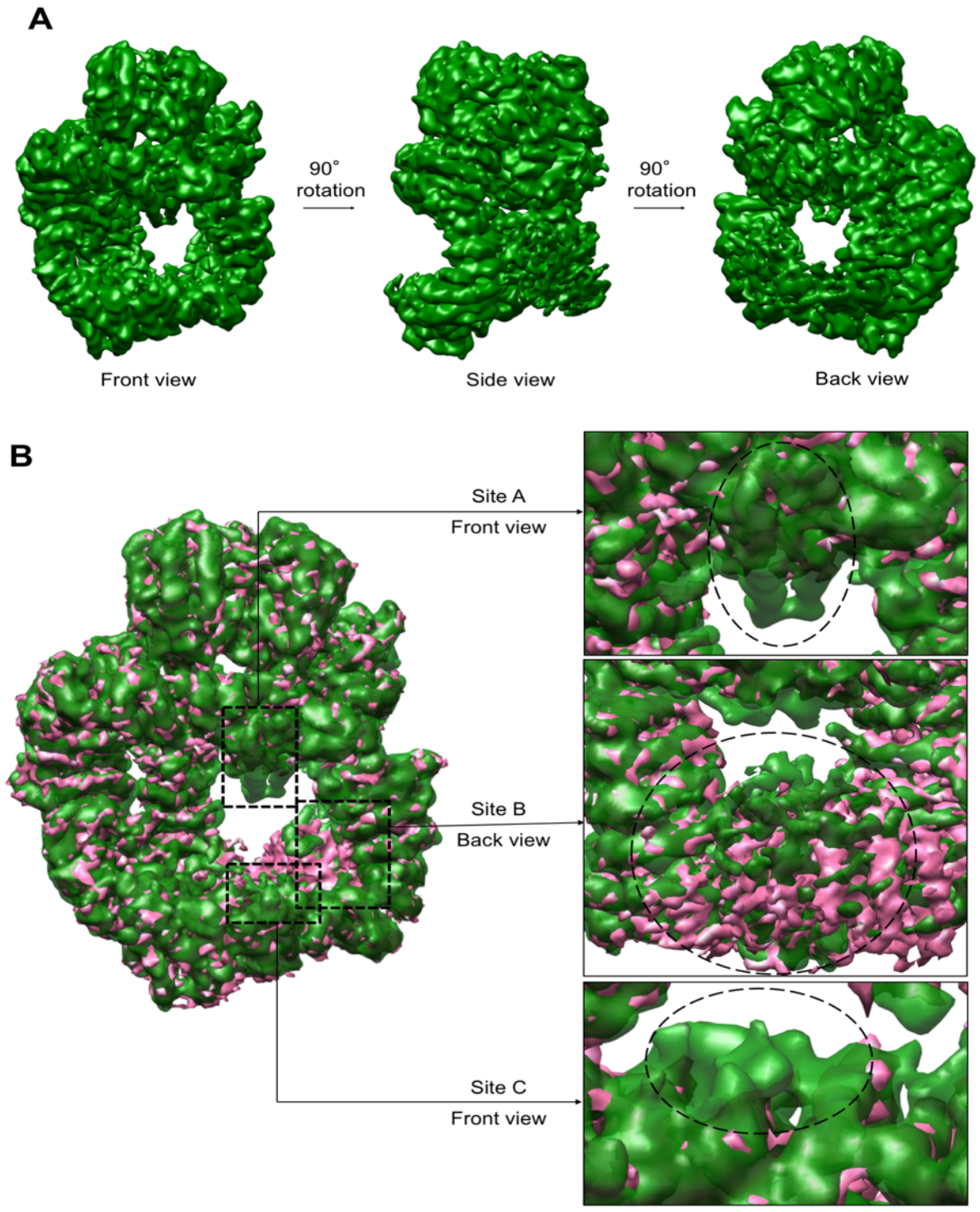


Figure 40. Analysis of the cryo-EM map at the resolution of 6.2 Å of the DNA-PKcs/ Artemis complex. (A) Overall presentation of the map with the front view, side view and back view. The most obvious feature of the map is the fragmented density of the N-terminal arm; (B) Comparison of the DNA-PKcs/ Artemis complex map and the published apo DNA-PKcs map (emd_8751). The DNA-PKcs/ Artemis complex map is coloured green while the apo DNA-PKcs map (emd_8751) is coloured pink. There are three regions that as shown in the figure. At site A, there is extra density that comes down from the circular cradle and hangs on top of the central cavity of DNA-PKcs. This is similar to what was observed in the X-ray crystallisation density map (PDB: 5LUQ). Site B is the N-terminal bridge region of DNA-PKcs. There are extra densities compared to the apo DNA-PKcs. However, unlike the apo form, the de-

-nsity of the N-terminal arm is highly fragmented and is more extensive than in the map of the complex, indicating that the full-length Artemis interacts with the N-terminal arm and results in conformational changes. Site C sits on the circular cradle of DNA-PKcs and there is a tube of extra density on top of the apo DNA-PKcs HEAT repeats.

Although it is difficult to fit the sequence to the map due to the low resolution, extra densities can be identified in the cryo-EM map of DNA-PKcs/ Artemis complex compared to that of apo DNA-PKcs. To have a clearer view of the DNA-PKcs/ Artemis interaction, further analyses were carried out and the sites A, B and C are highlighted to make clearer the changes related to the DNA-PKcs/ Artemis interaction.

6.3 Cryo-EM of DNA-PKcs/ Artemis/ DNA Complex

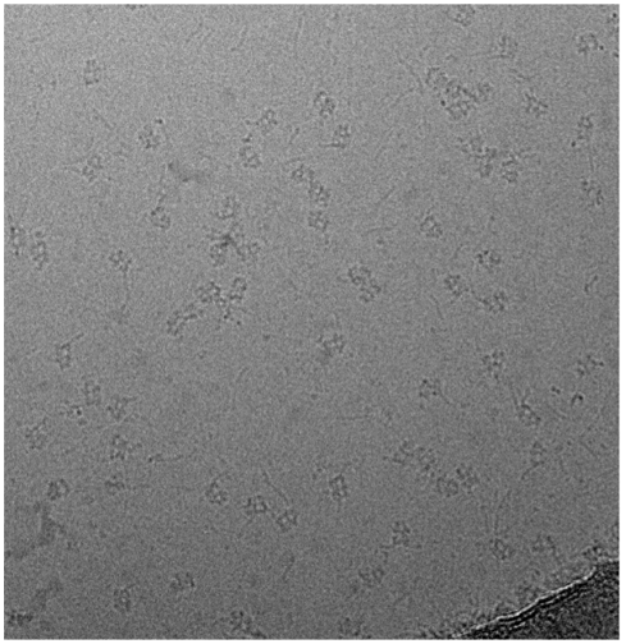
Cryo-EM was then used to study the structure of DNA-PKcs/ Artemis/ DNA complex to understand the interaction among all three components of the endonuclease complex. As the Artemis nuclease region was not detected in the previous DNA-PKcs/ Artemis complex, adding DNA may be able to stabilise and dock the nuclease region of Artemis to a certain conformation. To ensure that the Artemis will not cut the DNA and can be docked nicely on the DNA ends, Artemis H115A instead of wild-type Artemis was used.

6.3.1 Sample preparation, grid screening and data collection

The purified DNA-PKcs, at a final concentration of 0.5 mg/ml, was added to Artemis H115A and hpDNA in a 1:2:2 ratio. The final buffer was the same as the nuclease assay buffer and the hpDNA was the construct used for the nuclease assay without the fluorescence label. Vitrobot was used for the grid preparation and aliquots of 2.5 µl sample were applied to glow-discharged grids (Quantifoil Au R1.2/1.3, 300 mesh), which were then blotted for 3 s with filter paper and plunge-frozen in liquid ethane.

The grids were first screened and an example of the micrographs is shown in the figure (Figure 41). The micrographs appeared to have a better signal-to-noise ratio with a cleaner background and better contrast than those of the DNA-PKcs/ Artemis complexes (Figure 38). This could be due to the change of the grid material from copper to gold. In the micrograph, particles of DNA-PKcs can be clearly seen evenly distributed. However, the density of the particles is lower than that of the DNA-PKcs/ Artemis complex (Figure 38). In addition, the association of DNA-PKcs particles as long filaments does not occur. It may be caused by changes of the surface property of the complex due to the addition of DNA or by changes of the grid type. The hpDNA added to the system can be clearly visualised as short strings, some of which float in the buffer on their own while others are linked with the particles of DNA-PKcs. There are also many little black dots in the micrographs that are likely to be the structured nuclease region of Artemis. The black dots are distributed across the micrograph in different ways, either as individual single particles or close to the DNA-PKcs particles but not in contact.

The screened grid was used for data collection on the Thermo Fisher Titan Krios electron microscope (300kV) with a Gatan K2 detector and the movies were recorded under the counting mode using a magnification of 130,000 times. The pixel size used for the collection was 1.05 Å. Total dose for a movie stack was 60 electrons/Å^2 with an exposure time of 12 s. The same defocus value series were used for collection including -1.3, -1.6, -1.9, -2.2, -2.5, -2.8, and -3.1 µm. Eventually 2461 movies were collected and later processed.



Microscope	Titan Krios
Detector	K2 (counting)
Nominal Magnification	130,000x
Pixel Size, Å per pixel	1.05
Total Dose, electrons/Å^2	60
Exposure, sec	12
Defocus Range, µm	-1.3 ;-1.6; -1.9; -2.2; -2.5; -2.8; -3.1

Figure 41. Cryo-EM screening and data collection of DNA-PKcs/ Artemis/ DNA complex. Left panel is an example of the screened micrographs of the grids of DNA-PKcs/ Artemis/ DNA complex. There are evenly distributed DNA-PKcs particles in the micrographs. The hpDNAs added behave like strings. Some remain on their own while others appear connected to DNA-PKcs particles. The little black dots floating around the micrographs are likely the signals from the nuclease region of Artemis. Some of the dots are individual particles with no attachment to DNA-PKcs or the hpDNA strings while some of them remain very close to the particles of DNA-PKcs; Right panel shows the parameters used for the final data collection. The dataset was collected at the Cryo-EM facilities at the UK national electron bio-imaging centre (eBIC).

6.3.2 Data processing

Different procedures for data processing for DNA-PKcs/ Artemis/ DNA complex were explored. The best map achieved was at the resolution of 6.6 Å following the approach shown in Figure 42.

First, the 2461 micrographs collected were sent for motion correction and CTF estimation. 2158 micrographs were selected with good CTFs, used for the subsequent particle auto-picking. As the backgrounds of the micrographs are clean with good signal-to-noise ratio, the auto-picking behaved relatively well. Considering the low density of the particles, manual cleaning was done on all 2158 micrographs to remove all contamination and add the unpicked particles. 67852 particles were finally picked and sent to 2D classification. After 2D classification, particles that could not be categorised by the software and particles of some contamination were discarded, leaving 60300 particles for the following 3D classification. Those particles were separated into four classes. However, the software did not manage to identify the differences, resulting in four similar classes with a similar number of particles. Therefore, all the particles were combined together for 3D refinement and postprocess. At the end, a cryo-EM map of DNA-PKcs/ Artemis/ DNA complex at the resolution of 6.6 Å was obtained.

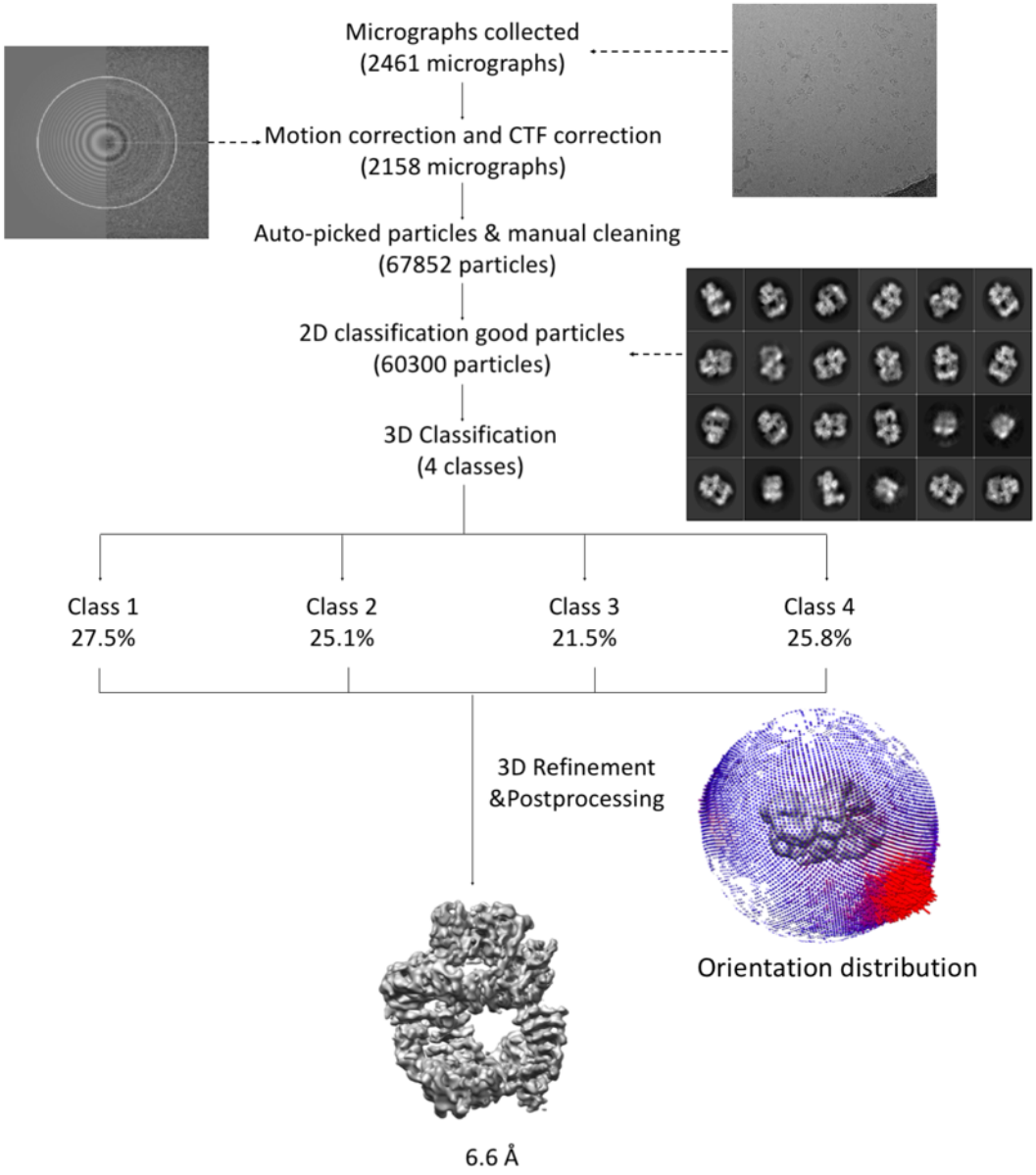


Figure 42. Cryo-EM data processing pathway of the map of DNA-PKcs/ Artemis/ DNA complex at the resolution of 6.6 Å. 2158 micrographs were picked from the original 2461 collected micrographs after motion correction and CTF estimation. 67852 particles were extracted from the micrographs after auto-picking and manual-cleaning. 2D classification was then applied to the extracted particles and 60300 were selected from them by removing particles that could not be categorised by the 2D classification and that have repeated views. 3D classification of the four classes of particles did not distinguish differences between the particles. Therefore, all four classes were combined together to proceed through 3D refinement and postprocessing to achieve a cryo-EM map of DNA-PKcs/ Artemis/ DNA complex at the resolution of 6.6 Å.

6.3.3 Cryo-EM map analysis

The map of the complex shows clearly the structure of DNA-PKcs with a protruding tail coming from the N-terminal arm (Figure 43A). There is extra density in the region of N-terminal arm that differs from the density of the complex of DNA-PKcs/ Artemis. Although the density at N-terminal arm is at low resolution, coming from weak signals that disappear first in the map when tuning down the contour level, the density is less fragmented. This indicates that, despite of a certain amount of flexibility, the N-terminal region here has a more fixed conformation compared to that of the DNA-PKcs/ Artemis complex (Figure 43B site B). Also, the helical protruding tail coming from DNA-PKcs is similar to the tail visualised in the map of DNA-PK when DNA is added to the system (Yin *et al.,* 2017). The exact position of the DNA/ DNA-PKcs interaction site and the level of the uplift of the N-terminal arm are not identical but very close to those observed in DNA-PK. Therefore, without Ku, DNA can still gain access close to the N-terminal arm of DNA-PKcs. Due to the low resolution it remains unclear as to whether any part of the C-terminal tail of Artemis is contributing to the extra density at site B but it is clear that the Artemis nuclease region is not docked at the DNA end here as it is not possible to fit a structured region of 40kDa into the extra density at site B. This is consistent with the observation of the micrographs that the Artemis nuclease region (black dots in the micrographs) are not likely to dock on DNA-PKcs together with the DNA. The interaction mode of DNA-PKcs and Artemis is very different from that of DNA-PKcs and Ku, where the globular domain of Ku comes into contact with the N-terminal arm of DNA-PKcs with the help of DNA.

In addition to the difference at the N-terminal region of DNA-PKcs (Figure 43B site B), comparison of this map with that of the DNA-PKcs/ Artemis complex also shows further differences at sites A and C (Figure 43B). In both sites the aforementioned extra densities in DNA-PKcs/ Artemis complex were all absent from the new map. It is obvious that DNA changes the binding interaction between Artemis C-terminal tail and DNA-PKcs.

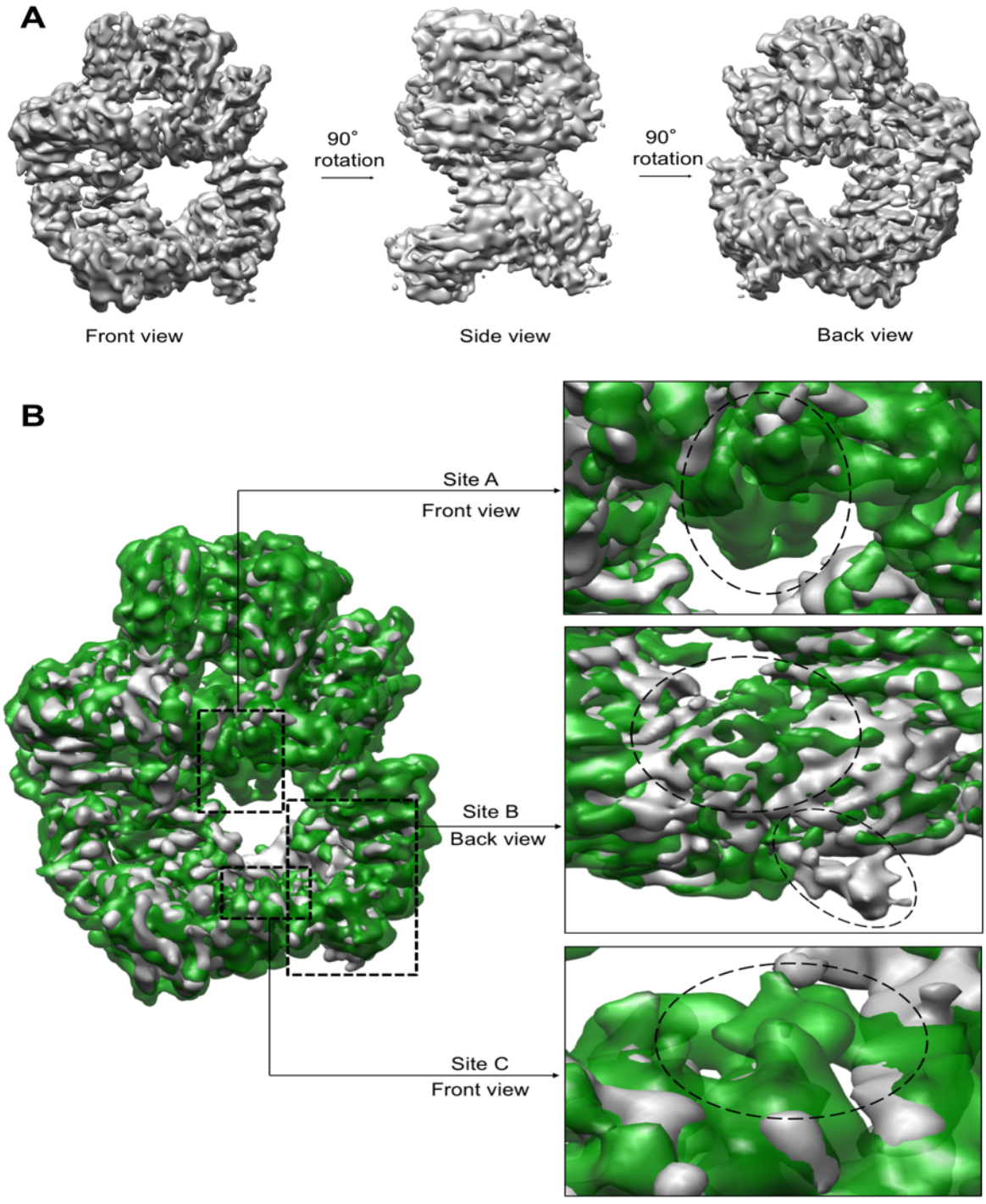


Figure 43. Analysis of DNA-PKcs/ Artemis/ DNA complex cryo-EM map at 6.6 Å resolution. (A) Overall presentation of the map with front, side and back views. The most obvious characterisation of the map is the protruding helical tail coming from the N-terminal arm of DNA-PKcs; (B) Comparison of the DNA-PKcs/ Artemis/ DNA complex map and the DNA-PKcs/ Artemis complex map. The map of DNA-PKcs/ Artemis complex is coloured grey, while that of DNA-PKcs/ Artemis complex is coloured green. Sites A, B and C, where differences between apo DNA-PKcs and DNA-PKcs/ Artemis complex were evident, are examined in the new map. The extra densities at sites A and C are not present in the map when DNA is present. In site B, the density become continuous and less uplifted.

6.4 Cryo-EM of DNA-PKcs/Artemis 399-426 Complex

Although the previous cryo-EM maps of DNA-PKcs/ Artemis and DNA-PKcs/ Artemis/ DNA showed the extra density for Artemis and DNA, it is difficult to define the interaction surface of DNA-PKcs and full-length Artemis due to the low resolution. This is likely to be due to the flexibility of both DNA-PKcs and the C-terminal tail of Artemis. Therefore, to reduce the level of flexibility and so improve the resolution, a cryo-EM study of the DNA-PKcs and Artemis 399-426 complex was conducted, following the demonstration described in section 5.1.3 that Artemis 399-426 interacts with DNA-PKcs.

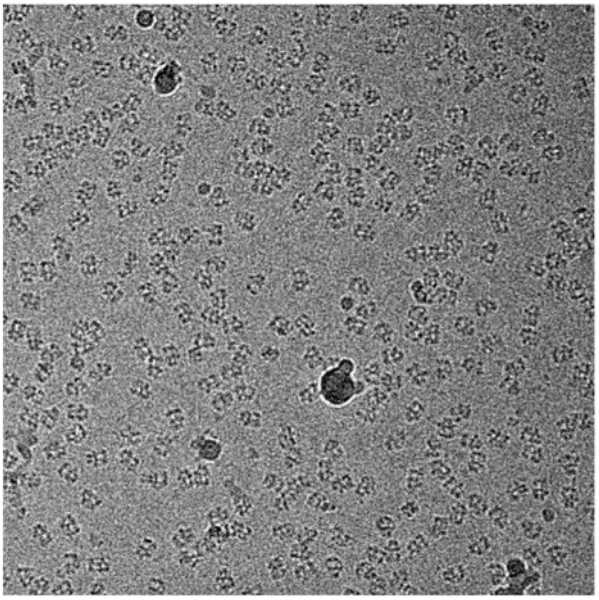
6.4.1 Sample preparation, grid screening and data collection

Purified DNA-PKcs, at a final concentration of 0.3 mg/ml, and the synthesised peptide of Artemis 399-426 were mixed at a 1:3 ratio. The final buffer was the same as that of the pull-down experiment. Vitrobot was used for the grid preparation and aliquots of 2.5 µl sample were applied to glow-discharged grids (Quantifoil Cu R1.2/1.3, 300 mesh), which were then blotted for 3 s with filter paper and plunge-frozen in liquid ethane.

The grids were first screened and an example of the micrographs is shown in the figure (Figure 44). The particles of DNA-PKcs are evenly distributed across the micrograph. Unlike the cases of DNA-PKcs/ Artemis and DNA-PKcs/ Artemis/ DNA, the particles of DNA-PKcs appear to be of high density and homogeneous, staying as individual particles without forming any filaments or complexes with others. As the Artemis peptide has only 28 amino acid residues, it is impossible to visualise it in the micrograph directly and high-resolution information is needed to detect the presence of the peptide. However, it is clear that adding the Artemis 399-426 peptide has a completely different effect on DNA-PKcs compared to full-length Artemis. It indicates that there may be further interactions between DNA-PKcs and full-length Artemis but higher-resolution details of the interactions are needed for further understanding.

The screened grid was applied to data collection on the Thermo Fisher Titan Krios electron microscope (300kV) with a Gatan K2 detector and the movies were recorded under the counting mode with a magnification of 130,000 times. The pixel size used for the collection

was 1.07 Å. Total dose for a movie stack was 63.8 electrons/Å^2 with an exposure time of 10 s. The defocus value series used for collection includes -1.0, -1.3, -1.6, -1.9, -2.2, -2.5, and -2.8 μm. Eventually 1747 movies were collected after the data collection and processed later.



Microscope	Titan Krios
Detector	K2 (counting)
Nominal Magnification	130,000x
Pixel Size, Å per pixel	1.07
Total Dose, electrons/Å^2	63.8
Exposure, sec	10
Defocus Range, μm	-1.0; -1.3; -1.6; -1.9; -2.2; -2.5; -2.8

Figure 44. Cryo-EM screening and data collection of DNA-PKcs/Artemis 399-426 complex. The left panel is an example of the screened micrographs of the grids of DNA-PKcs/Artemis 399-426 complex. The particles of DNA-PKcs are nicely distributed across the micrograph as individual particles, although Artemis 399-426 peptide cannot be visualised in the micrograph directly. The right panel shows the parameters that were used for the final data collection. The dataset was collected at the cryo-EM facility of the Department of Biochemistry, University of Cambridge.

6.4.2 Data processing

Various procedures for data processing for DNA-PKcs/ Artemis 399-426 were explored. As predicted, decreasing the flexibility of Artemis improved the resolution considerably and the best map achieved was at 4.2 Å resolution using the procedure shown in Figure 45.

First, the 1747 micrographs finally collected were sent for motion correction and CTF estimation. 1670 micrographs with good CTFs were selected for subsequent particle auto-picking. Thanks to the high density of the particles, 500180 particles were finally auto-picked and sent to 2D classification. After 2D classification, contamination and bad classes were discarded, leaving 374087 particles for later 3D classification. Those particles were separated into eight classes. All 3D maps were compared and the two classes with similar extra density compared to apo DNA-PKcs (class1 and class8). They were then combined for further 3D refinement and postprocessing and a cryo-EM map of DNA-PKcs/ Artemis 399-426 complex at a resolution of 4.9 Å was obtained, which was of better resolution compared to the previous cryo-EM maps. To further polish the map, Bayesian polishing, a new function of Relion 3.0, was used to reconstruct the map of 4.9 Å. After polishing, 3D refinement and postprocess were reapplied and the final map of DNA-PKcs/ Artemis 399-426 complex at a resolution of 4.2 Å was obtained.

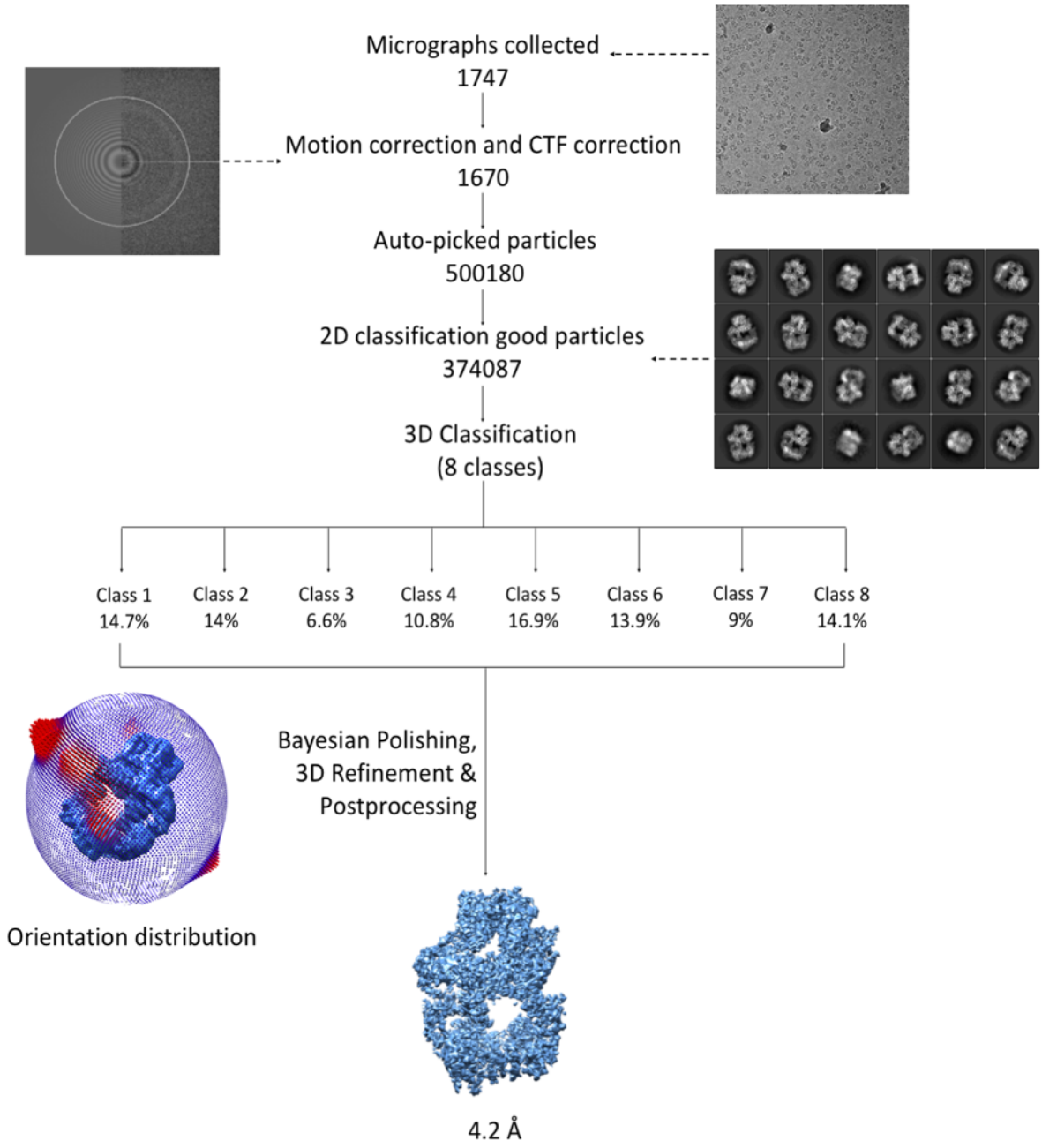


Figure 45. Cryo-EM data processing procedure for the map of the DNA-PKcs/ Artemis 399-426 complex at a resolution of 4.2 Å. 1670 micrographs were picked from the 1747 collected micrographs after motion correction and CTF estimation. 500180 particles were auto-picked and extracted from the micrographs. 2D classification was then applied to the extracted particles and 374087 particles were selected after filtering. 3D classification for eight classes of those particles was able to distinguish differences. The two classes (class 1 and class 8) with similar extra densities were combined together to go through 3D refinement and postprocessing to achieve a cryo-EM map of DNA-PKcs/ Artemis 399-426 complex at the resolution of 4.9 Å. Later, Bayesian polishing was applied to the particles and a second round of 3D refinement and postprocessing on the polished particles produced a new map of DNA-PKcs/ Artemis 399-426 complex at the resolution of 4.2 Å.

6.4.3 Cryo-EM map analysis

Generally, the map of DNA-PKcs/ Artemis 399-426 complex is more similar to that of apo DNA-PKcs than that of the DNA-PKcs/ Artemis complex (Figure 46A).

Comparison between the maps of DNA-PKcs/ Artemis 399-426 complex and DNA-PKcs/ Artemis complex allowed further insights into the structure (Figure 46B). The extra density in the DNA-PKcs/ Artemis 399-426 map is mainly located on site C (Figure 46B site C). This extra density is also present in the map of DNA-PKcs/ Artemis complex, indicating that it belongs to the sequence of 399-426. However, the previously observed extra density in site A, which is likely to be part of DNA-PKcs, in the DNA-PKcs/ Artemis complex is now gone. In site B, which is also one of the full-length Artemis interaction sites, the extra fragmented density that is likely to come from Artemis C-terminal tail peptides is also absent. The similarities and differences of the maps indicate that the interaction modes of DNA-PKcs/ Artemis and DNA-PKcs/ Artemis 399-426 are different. Artemis 399-426 is likely to be involved in the interaction of full-length Artemis and DNA-PKcs, but there should be other regions of the Artemis C-terminal tail interacting/ interfering with the N-terminal bridge of DNA-PKcs.

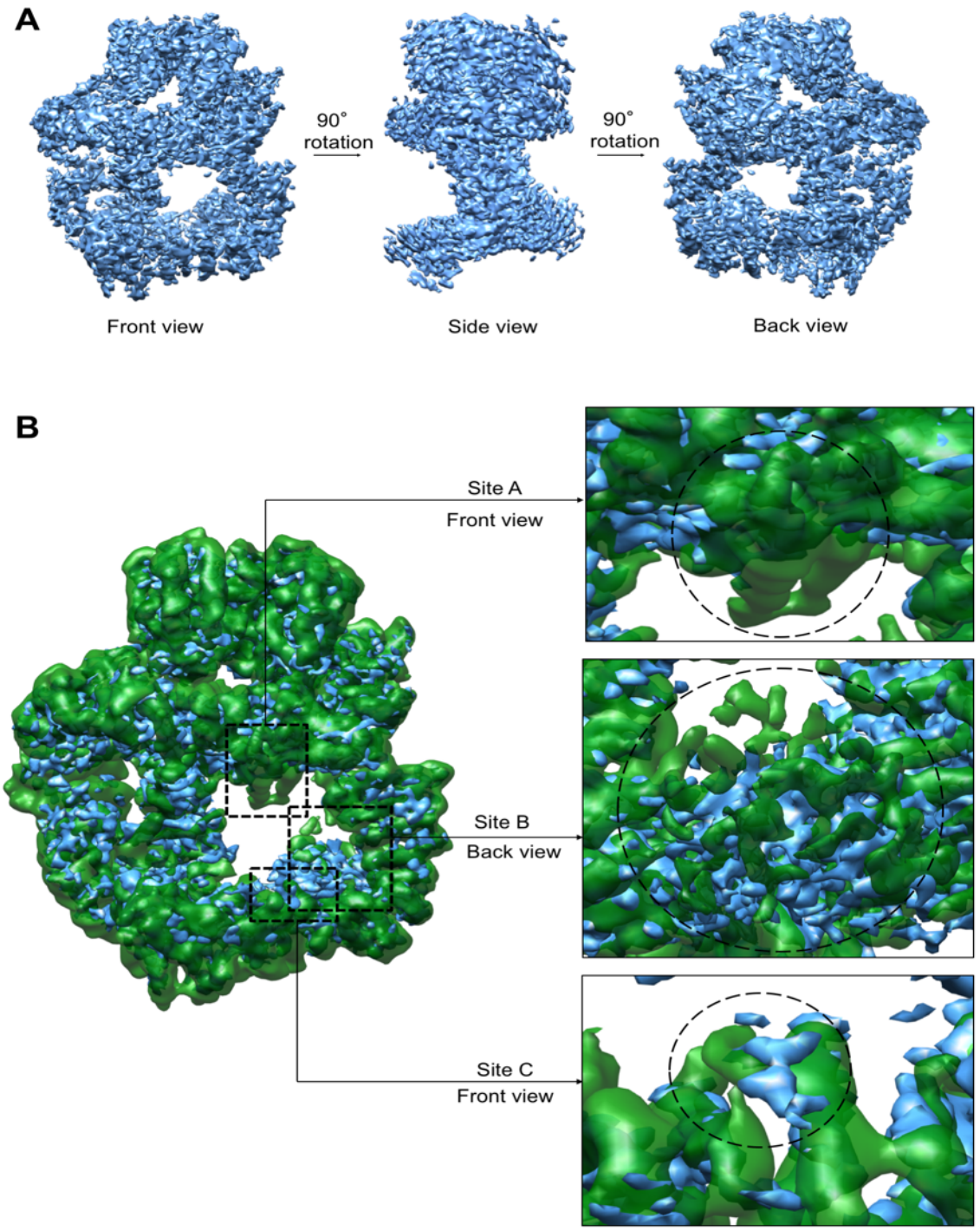


Figure 46. Analysis of DNA-PKcs/ Artemis 399-426 complex cryo-EM map at the resolution of 4.2 Å. (A) Overall presentation of the map with the front view, side view and back view; (B) Comparison of the DNA-PKcs/ Artemis 399-426 complex map and the DNA-PKcs/ Artemis complex map. The DNA-PKcs/ Artemis 399-426 complex map is coloured blue while the DNA-PKcs/ Artemis complex map is coloured green. Site A, B and C, where differences showed up between apo DNA-PKcs and DNA-PKcs/ Artemis complex, are examined in the new map. The extra densities at site A is gone in the new map. As for site B, the volume of the density is smaller, more continuous and less uplifted. The extra density showed in site C still sits at the same place in the map of DNA-PKcs/ Artemis 399-426 complex

Moreover, detailed comparison between the maps of DNA-PKcs/ Artemis 399-426 complex and apo DNA-PKcs reveals more about the interaction pattern and provides a better picture of the extra density of Artemis 399-426.

Generally, the extra density of Artemis 399-426 is composed of two parts, the discontinuous linking region and the continuous binding region. As shown (Figure 47), there are many fragmented densities across the "hole" at the base of the circular cradle of DNA-PKcs. Although they are not continuous and it is difficult to define the sequence, there is a clear trend of the density linking regions of the circular cradle (residues 2080-2228) and the N-terminal HEAT repeat (257-305). While the fragmented densities are difficult to trace, probably due to flexibility of some parts of the Artemis peptide, there is a continuous region of density sitting on the circular cradle of DNA-PKcs, interacting with two α helices (residues 2176-2198 and 2230-2247) of the HEAT repeats. According to the cryo-EM structure of DNA-PKcs, the exposed surface of those two helices is highly negatively charged. Furthermore, considering the conservation of the positively charged residues in the Artemis peptide (402-404) and in the larger region of Artemis 399-426, and also previous studies on the importance of R402 to the Artemis/DNA-PKcs interaction, the N-terminal region of Artemis 399-426, including residue 402, is likely to be the continuous binding region. Unlike the Artemis/DNA Ligase IV interaction, in which the sequence of residues 485-495 within the generally intrinsically disordered region of Artemis undergoes concerted folding into a well-defined groove with regions to accommodate conserved aromatic sidechains of Artemis (Ochi et al.,), the interaction between Artemis and DNA-PKcs appears to be more of an unstructured peptide of Artemis sitting on the surface of DNA-PKcs. Based on the length of the density, 8 residues of Artemis are likely involved. Also, E2188, E2243, E2236 of DNA-PKcs, which are conserved among different species down to zebrafish, are very likely to be the key acidic residues to bind the basic residues R402, H403, K404 of Artemis.

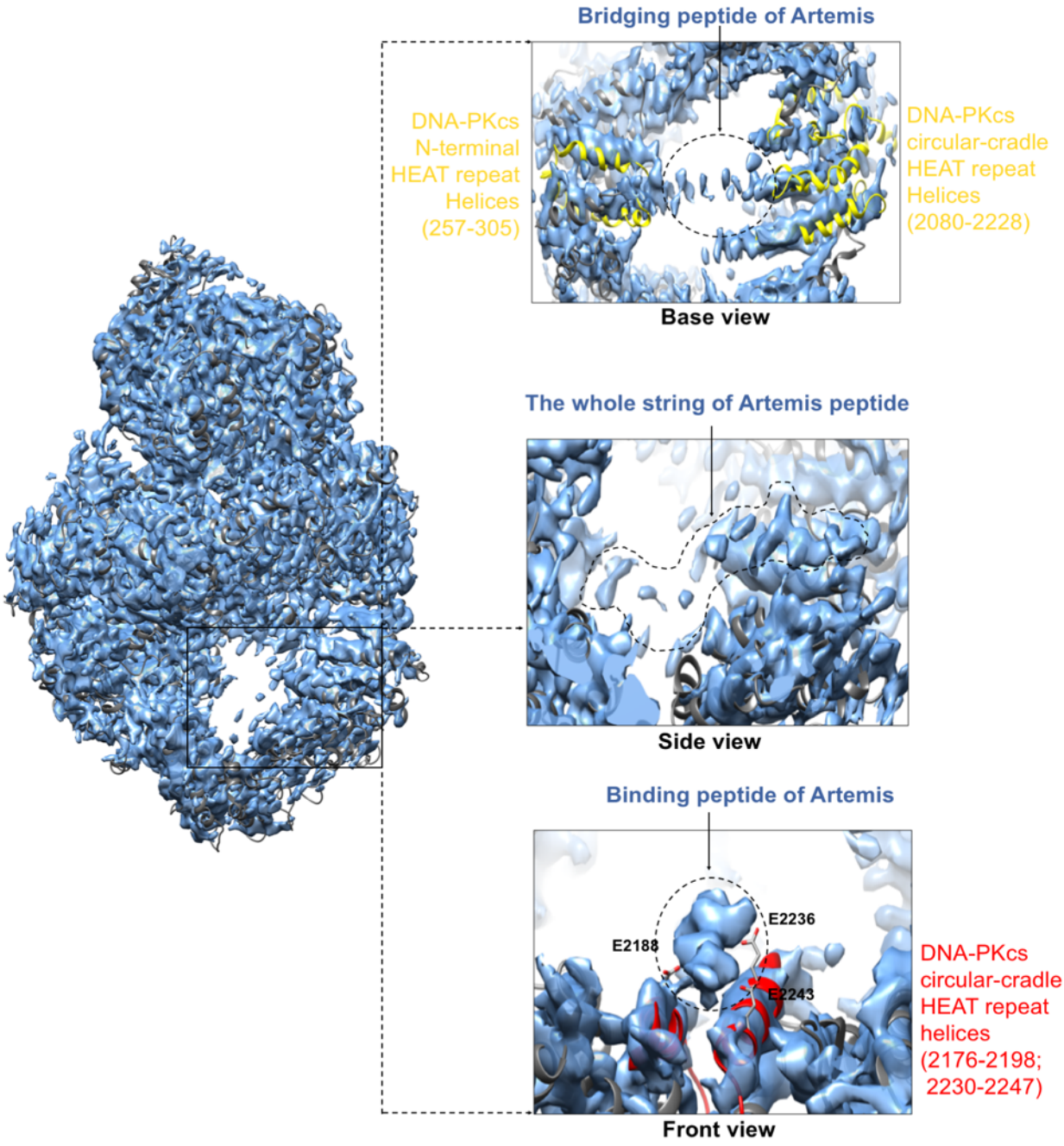


Figure 47. Cryo-EM map of DNA-PKcs/Artemis 399-426 complex in comparison with cryo-EM apo-DNA-PKcs model (PDB: 5W1R). The map of DNA-PKcs/Artemis 399-426 complex is coloured blue and apo-DNA-PKcs model (PDB: 5W1R) is coloured grey. From the side view panel, it is visualised that the extra density of Artemis 399-426 is like a string going through the circular cradle and the empty space at the base of DNA-PKcs and get in contact with the N-terminal HEAT repeats. The extra densities can be separated into two parts—the binding peptide, shown in the front view panel, and the bridging peptide, shown in the base view panel. The region where the binding peptide sits is highly negative with the E2188, E2236 and E2243 of DNA-PKcs, indicating that the binding peptide is likely to be the positive N-terminal region of Artemis 399-426. The bridging peptide goes across the empty space at the base of DNA-PKcs. It is clearly shown in the base view panel that the bridging peptide links the N-terminal HEAT repeat helices and the circular cradle HEAT repeat helices and closes the empty space.

Interestingly, the Artemis continuous binding region (~399-407) sits on the site where the extra density of Ku80 interacts with DNA-PKcs, which was proposed in both crystal structure of DNA-PKcs in complex with Ku80 CTD and cryo-EM structure of DNA-PK. In the crystal structure, two helices of Ku80 CTD, proposed to be between the globular region of K80 CTD and the C-terminal helix (595-732), bind on the platform of four helices (2177-2198; 2230-2247; 2267-2283; 2313-2334). In the cryo-EM structure, only one 4-turn helix can be accommodated in the density to interact with two helices of DNA-PKcs (2178-2197; 2231-2248). This helix is predicted to be one of the Ku80 CTD helices of the crystal structure based on the length (Yin et al., 2017). Although the models proposed for Artemis and Ku80 C-terminus differ and the sequences involved remain uncertain, Artemis C-terminal region is predicted in this model to compete for DNA-PKcs binding on the circular cradle. To test whether Artemis competes with Ku80 CTD, a serial pulldown of DNA-PKcs and Artemis 399-426 was conducted under the low-to-high concentration gradient of Ku80ct140 (593-732), which contains the region proposed to have the extra density of Ku helixes on the circular cradle. There was no inhibition at any concentration of the Ku80CTD peptide, indicating that Ku80CTD on its own is not enough to stop the interaction between Artemis and DNA-PKcs (Figure S3). Moreover, there may be other unknown interactions within the Ku/ DNA-PKcs/ Artemis system and a complex that accommodates all three proteins.

In general, the Artemis peptide 399-426 appears to be like a string, interacting with both the circular cradle and the N-terminal HEAT repeats (Figure 47). The interaction between Artemis and DNA-PKcs is not an interaction of an isolated single site but that of multiple sites in a connected way. The positive residues of Artemis around R402 interacting with the circular cradle HEAT repeats of DNA-PKcs are important and conserved. Moreover, the interaction between the C-terminal region of Artemis 399-426 (mainly Artemis 413-426) and DNA-PKcs N-terminal HEAT repeat further strengthens the binding. This indicates that Artemis is able to interact with DNA-PKcs in the presence of Ku and there may be a protein complex of higher magnitude.

Comparison of the model/map of apo DNA-PKcs and that of DNA-PKcs in complex with Artemis peptide also indicates that DNA-PKcs undergoes a conformational change (Figure 47 Base view). Although there was no evidence that Artemis has an allosteric effect on DNA-PKcs,

it shows that all the different domains of the DNA-PKcs molecule are connected and even slight changes of a small region could have allosteric effect on the whole molecule. One of the obvious conformational changes when Artemis C-terminal peptide interacts with DNA-PKcs is the shrinking of the base window. As the string of Artemis holds it closer together, both the circular cradle and the N-terminal HEAT repeats tend to close the window—especially N- terminal HEAT repeat region of 289-330 and circular-cradle HEAT repeat region of 2209-2246.

6.5 Cryo-EM of apo DNA-PKcs

Although various groups have been working on cryo-EM of DNA-PKcs and related complexes, none of the published maps is at high-resolution (better than 3.5 Å) (Sharif *et al.,* 2017; Yin *et al.,* 2017; Wu *et al.,* 2018). As a result, the modelling of the cryo-EM structures was based on the Blundell Group X-ray crystal structure at 4.3 Å, which used selenomethionine labelling for phase calculation and was a structure of DNA-PKcs bound to the C-terminus of Ku80. As there are various differences of the molecule between the crystal structure and cryo-EM structures, largely due to the nature of the complex, but also resulting from resolution and crystal packing, it is of great importance to define the apo-structure at higher resolution using cryo-EM. This may also allow *ab-initio* modelling to resolve some of the ambiguities in the sequence registration of the X-ray map, so providing firmer ground for the further study of DNA-PKcs related complexes.

As the field of cryo-EM has been developing very fast and new methods of sample preparation, data collection and data processing are evolving rapidly, I have been trying various strategies to define the cryo-EM structure of apo-DNA-PKcs at higher resolution. During the last months of experiments and analysis of my PhD study, I managed to obtain a map of apo DNA-PKcs with the highest resolution (compared to the published ones so far) and I briefly present the results here, although further work is required to model the complete sequence.

My apo-DNA-PKcs map has an overall resolution of 3.88 Å, with local resolution of the stable regions in the molecule of 3.3 Å (Figure 48). The kinase and nearby region sit in the heart of the "head" of the molecule and have high local resolution of around 3.3 Å. Unsurprisingly, the N-terminal arm has the highest flexibility and lowest local resolution. The other peripheral regions of the molecule also tend to be more flexible than the core region and exhibit lower resolution. This could be due to the biochemical properties of the hydrophilic residues on the interface as well as the data analysis algorithm. Therefore, sample optimisation and data processing optimisation may further improve the resolution. So far, the best map has been obtained from the recently introduced package-- cryoSPARC. Relion has also been used but, as yet, many other options of data processing including iterative particle polishing and multi-body refinement have not been tested and may lead to better maps. In general, this is the

best map of apo DNA-PKcs so far, indicating a possibility of achieving high resolution in the cases of DNA-PKcs and related complexes.

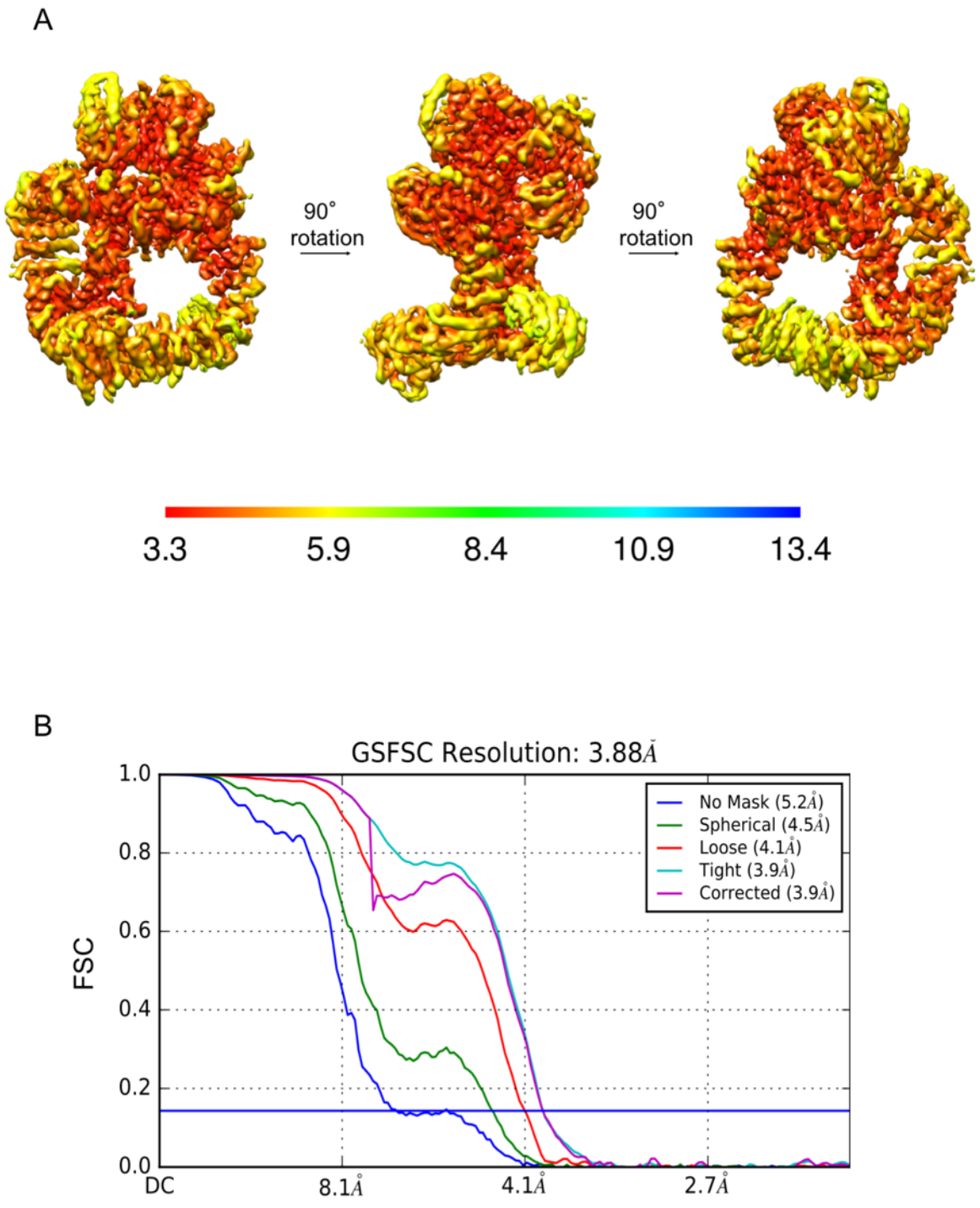


Figure 48. Cryo-EM rediscovery of DNA-PKcs. (A) Overall presentation of the DNA-PKcs map with the front view, side view and back view with coloured overall resolution distribution; (B) FSC curves of the final refined DNA-PKcs structure.

6.6 Summary

Although the resolution is limited, cryo-EM studies of the DNA-PKcs/ Artemis complex, DNA-PKcs/ Artemis/ DNA complex and DNA-PKcs/ Artemis 399-426 complex together provide insights for the first time of the complicated and flexible interactions between DNA-PKcs and Artemis:

First of all, although it is difficult to register the sequence of Artemis to the map, the map of DNA-PKcs/ Artemis 399-426 complex at 4.2 Å reveals the interaction between DNA-PKcs and Artemis 399-426. Artemis 399-426 interacts mainly with the circular cradle HEAT repeats (DNA-PKcs 2176-2198 and 2230-2247) and HEAT repeat in the N terminal region (DNA-PKcs 257-305). Moreover, the addition of Artemis 399-426 changes the conformation of DNA-PKcs molecule and partially closes the empty window at the base of the molecule as it pulls the N-terminal HEAT repeat and the circular cradle HEAT repeat closer.

Moreover, the map of DNA-PKcs/ Artemis complex demonstrates another mode of interaction. In addition to the density on the circular cradle HEAT repeats (DNA-PKcs 2176-2198 and 2230-2247), there is extra fragmented density, very likely also from Artemis C-terminal tail, at the site of the DNA-PKcs N-terminal arm. It indicates that there may be another previously unknown region of Artemis interacting with DNA-PKcs and moving the N-terminal arm to a more flexible status. This interaction may also stabilise/change the conformation of DNA-PKcs as there is extra density around the region of 2576-2744 of DNA-PKcs, which covers the ABCDE cluster.

Furthermore, when DNA is added to the system of DNA-PKcs and Artemis, the interaction between DNA-PKcs and Artemis changes. The extra density on the circular cradle HEAT repeats and that at the N-terminal arm no longer exist. Instead, a helical protruding tail, which is very likely to be DNA, appears from near the N-terminal arm, revealing that DNA and the full-length Artemis share similar binding site on DNA-PKcs.

Last but not least, my very recent cryo-EM study of DNA-PKcs proves that achieving near atomic resolution for it and related complexes is practical and *ab initio* modelling of the apo DNA-PKcs molecule can be improved to provide a firmer ground for understanding other complexes.

Chapter 7. Conclusion and Perspective

NHEJ is the preferred DNA DSB-repair pathway in humans. It is a highly dynamic and open process in which various proteins and complexes may get involved, depending on the nature of the DNA damage, at different stages from DNA end recognition, DNA end synapsis and processing to the final DNA end ligation. Among the numerous components, DNA-PKcs and Artemis interactions are the central focus of my PhD project. DNA-PKcs/ Artemis is the main endonuclease involved in the DNA end-processing step. It is also the only endonuclease discovered in humans capable of cleaving DNA hairpins, which is indispensable in V(D)J recombination, which provides immunodiversity in antibodies and T-cell receptors. Despite its importance, little is known about the interaction pattern between DNA-PKcs and Artemis. Unlike most of the main NHEJ components with their full or major structures solved (Ku, DNA-PKcs, XRCC4, XLF, PAXX, DNA ligase IV), no structural information of any Artemis or related complexes is available except for the Artemis 485-495 peptide in complex with DNA ligase IV catalytic domain (Ochi, Gu and Blundell, 2013). So far, published work regarding the DNA-PKcs/ Artemis interaction mechanism has demonstrated that two residues of Artemis (L401 R402) are important but the extent of the region of Artemis binding to DNA-PKcs remains unknown, not to mention the interaction region of DNA-PKcs (Soubeyrand *et al.*, 2006). During my PhD, the mode of interaction between the two proteins was investigated and revealed for the first time.

First, my research has shown that Artemis peptide (399-426), including residues of L401 and R402, was identified as sufficient for the interaction between DNA-PKcs and Artemis. This was achieved by designing different peptides covering L401 and R402 for a series of pulldown experiments on the basis of a bioinformatics analysis. Secondly cryo-EM of DNA-PKcs/ Artemis 399-426 complex further defined the interaction pattern. There are two main binding sites of the peptide with DNA-PKcs—one on the circular cradle HEAT repeat (2176-2198 and 2230-2247) and the other at the HEAT repeat involving residues 257-305 in the N-terminal region. The peptide acts like a string, crossing the empty window at the base of DNA-PKcs to link the two interaction sites. The cryo-EM map also proves that the two regions around 401-402 and 413-426 are both important for the interaction and explains why neither region is

able to pull down DNA-PKcs on its own. Moreover, the linkage by the peptide induces conformation change in DNA-PKcs, mainly contracting the size of the base window but also leading to other movements within the whole molecule that probably have mechanistic consequences.

Cryo-EM work on the DNA-PKcs/ full length Artemis complex shows that further interactions between the two proteins may be important. In addition to the extra density shown in the DNA-PKcs/ Artemis 399-426 complex, there is other extra density at the N-terminal arm of DNA-PKcs. As the extra density, close to the FAT region, is fragmented, it may arise from the C-terminal tail of Artemis rather than the nuclease region. Moreover, as the density of the N-terminal arm (especially the HEAT repeat region of 8-167) becomes less ordered, this specific interaction of Artemis may interfere with the docking of N-terminal arm with the circular cradle and may 'open up' the region. Besides, Ku and DNA bind to the same site based on the DNA-PK structure, indicating potential competition between them and Artemis.

When DNA is added to DNA-PKcs/ Artemis complex, the protein-protein interaction mode changes. This could be caused by the competition between DNA and Artemis at the same binding site. Based on the map, DNA occupies the same N-terminal interaction site of DNA-PKcs, forming a protruding "tail". Moreover, unlike the case of DNA-PK, there was no extra density of Artemis nuclease region (globular region) docking around DNA-PKcs.

The different interaction pattern may underlie the mechanism of the endonuclease activation. Indeed previous research indicated that the region around Artemis 456-458 is important for self-interaction and auto-inhibition of endonuclease activity (Niewolik *et al.*, 2017). This is consistent with Artemis C-terminal tail interaction with DNA-PKcs inhibiting intramolecular interactions and reversing auto-inhibition. Considering the similar DNA-PKcs binding site between Artemis C-terminal tail and DNA, the endonuclease complex is likely to be activated in a trans fashion, meaning that the DNA molecule necessary for activation is not the one to be cleaved by the activated endonuclease.

To know more about the endonuclease mechanism, it will be useful to solve the structure of the Artemis nuclease region. Previous attempts at crystallisation of the full-length molecule

were unsuccessful, probably due to high flexibility of low complexity regions of the tail and low stability of the protein. However, isolation of the nuclease region alone has not proved successful, indicating possible stabilising interactions between the nuclease region and the tail. Although Artemis is a relatively small protein, with only about half of it forming the nuclease region, cryo-EM may also be useful later. With the development of cryo-EM, it may be highly challenging but practical to study the structured region of Artemis. Recently, cryo-EM structure of the 52 kDa streptavidin showed the potential of resolving small proteins under the help of Volta Phase Plate (VPP) (Fan *et al.,* 2019).

Not only the structure but also the function of DNA-PKcs/ Artemis endonuclease complex was investigated during my PhD, revealing the effect of other NHEJ components. Ku slowed down the endonuclease activity. This is consistent with the structural study as Ku interferes the interaction between DNA-PKcs and Artemis on both the N-terminal region and the circular cradle region. Surprisingly, XLF at high concentration strongly stimulates the endonuclease activity. XLF interacts with neither DNA-PKcs nor Artemis. Therefore, this stimulation should be caused by the interaction between XLF and DNA. Considering the trans activation of the endonuclease complex, it is likely that high-concentration XLF binds to DNA to enrich the local substrate concentration for the endonuclease interaction. Moreover, clinical data also show that there are mutations of four NHEJ component genes among patients with radiosensitivity and severe combined immunodeficiency (RS-SCID) including DNA-PKcs, Artemis, DNA Ligase IV and XLF (Woodbine, Gennery and Jeggo, 2014), suggesting that XLF likely plays a role in the endonuclease activity. However, when Ku is involved in the *in vitro* nuclease assay, the stimulation by XLF was inhibited. The reason could be that there was interference of the XLF-DNA interaction due to the strong binding between Ku and DNA. Nevertheless, high-concentration XLF/XRCC4 complex can reduce the interference of Ku and increase the endonuclease activity. Though the XLF/XRCC4 complex was well studied, this filament may have undiscovered functions in DNA-PKcs/Artemis endonuclease activity.

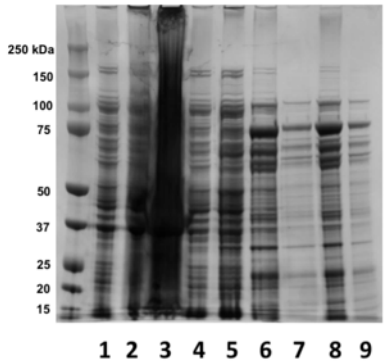
In addition to the endonuclease complex, the interaction of Artemis with DNA ligase IV was studied as well during my PhD. Previous work from our group revealed the structure of DNA Ligase IV/ Artemis 485-495 complex (Ochi, Gu and Blundell, 2013), indicating a good site for

drug molecule development. Therefore, initial FBDD was carried out to assess whether good fragment hits could be obtained that would allow further elaboration to drug-like molecules. Furthermore, the temporal organisation of NHEJ was studied through collaboration with the Strick group in Paris during my PhD. Using novel DNA forceps in a single-molecule approach, we were able to track the temporal organisation of NHEJ for the first time. The full NHEJ process was successfully recapitulated. More importantly, we discovered the importance of PAXX and its role in early DNA end synapsis. Using this platform, we may be able to reveal the time points when other unknown NHEJ components participate in the process.

With the development of cryo-EM, there are now things that can be done in addition to those previously envisaged. For instance, with the improvement on the methods for classification in data analysis, it is possible to conduct time-resolved structural studies. This can capture structures of transient states and enzymatic reactions, helpful in understanding many interaction processes within NHEJ. For example, the stepwise interactions between DNA-PKcs and Ku might be investigated to understand the transition from Ku80 CTD to the full Ku70/80 complex with DNA-PKcs and DNA. Furthermore, the improvement of data analysis may provide us with powerful in silico purification, which may be able to distinguish previously unknown complexes involved in the dynamic and flexible NHEJ. Last but not least, the improvement of cryo-ET and site-specific cryo-Focused Ion-Beam (FIB) sample preparation makes *in situ* structural study possible, in which case we should be able to investigate how NHEJ develops on damaged DNA ends in cell. Altogether I am looking forward to the future of NHEJ structural studies using cryo-EM!

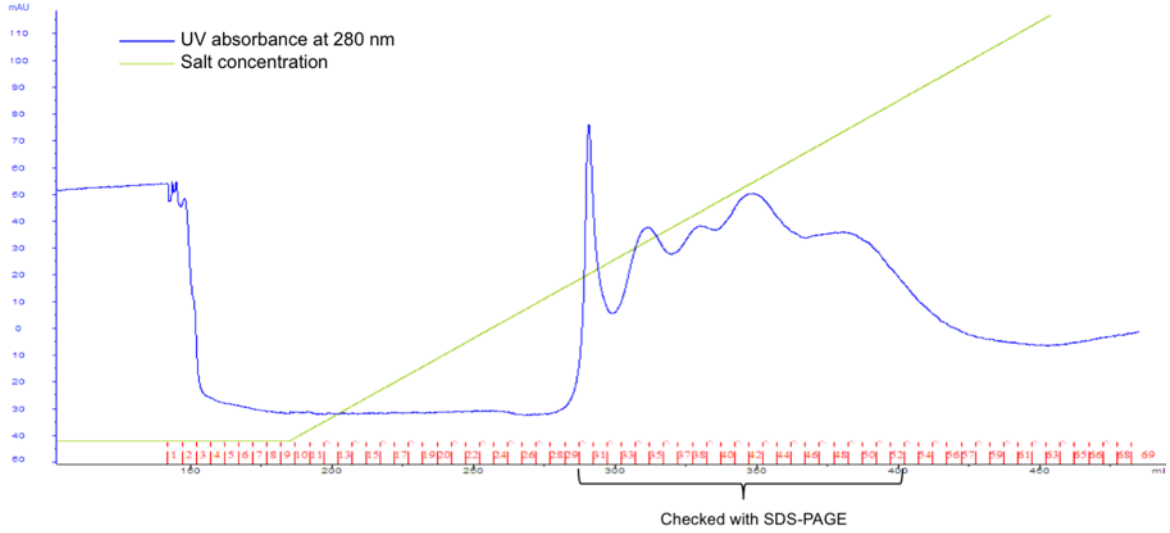
Supplementary Data

A

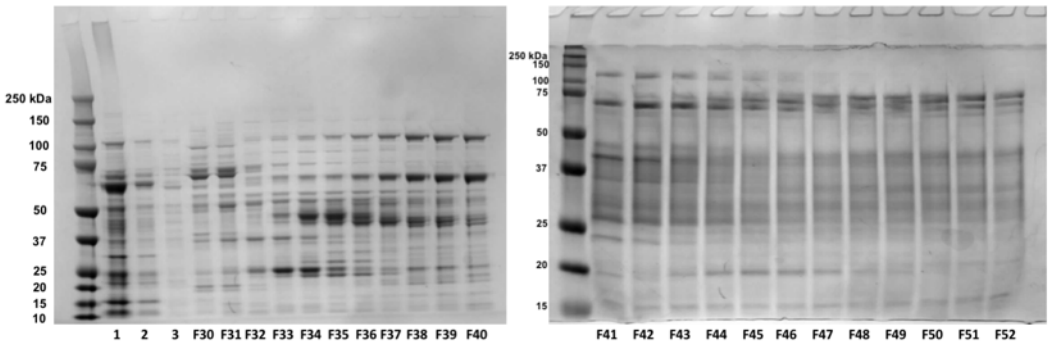


1: Cell lysis
2: Supernatant
3: Pellet
4: Flowthrough
5: Wash
6: Elution 1
7: Elution 2
8: Overnight dialysed elution 1
9: Overnight dialysed elution 2

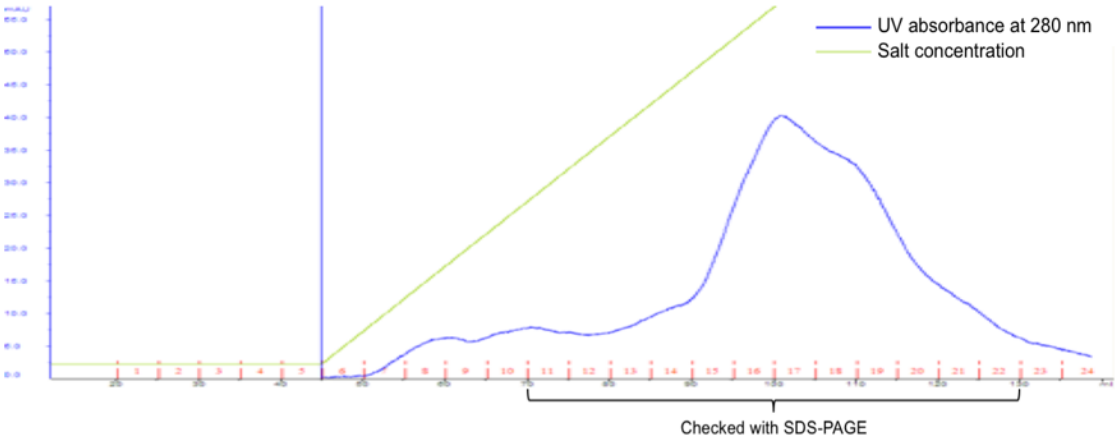
B

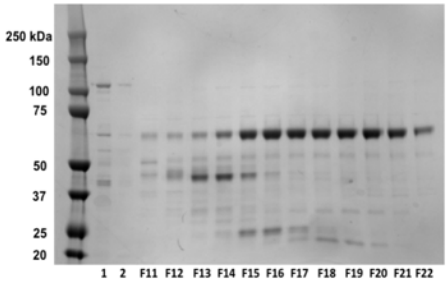


Checked with SDS-PAGE



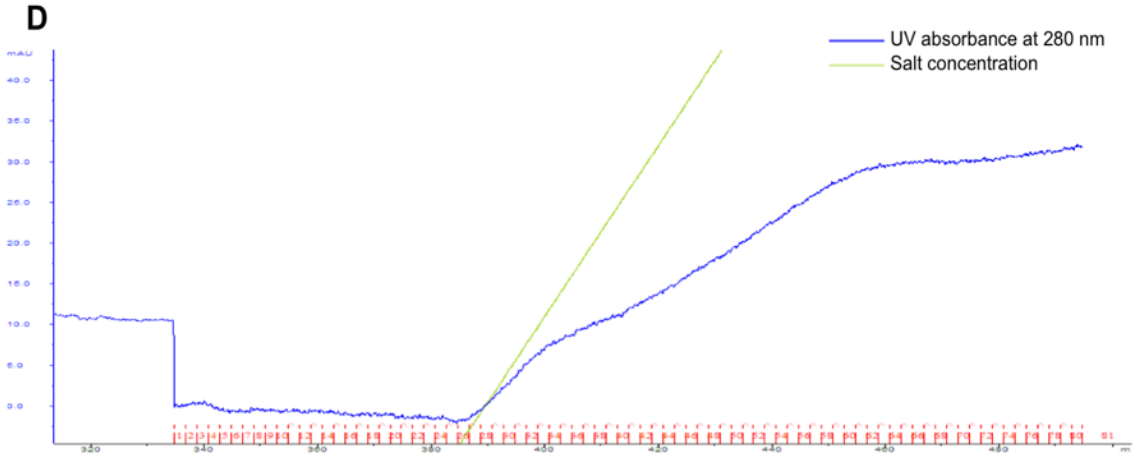
C

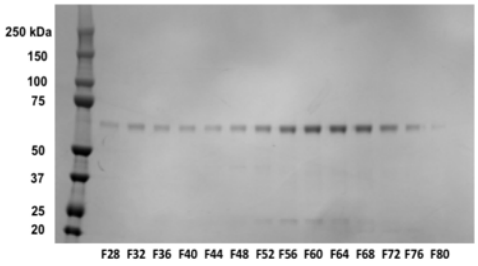




1: Flowthrough
2: Wash

D





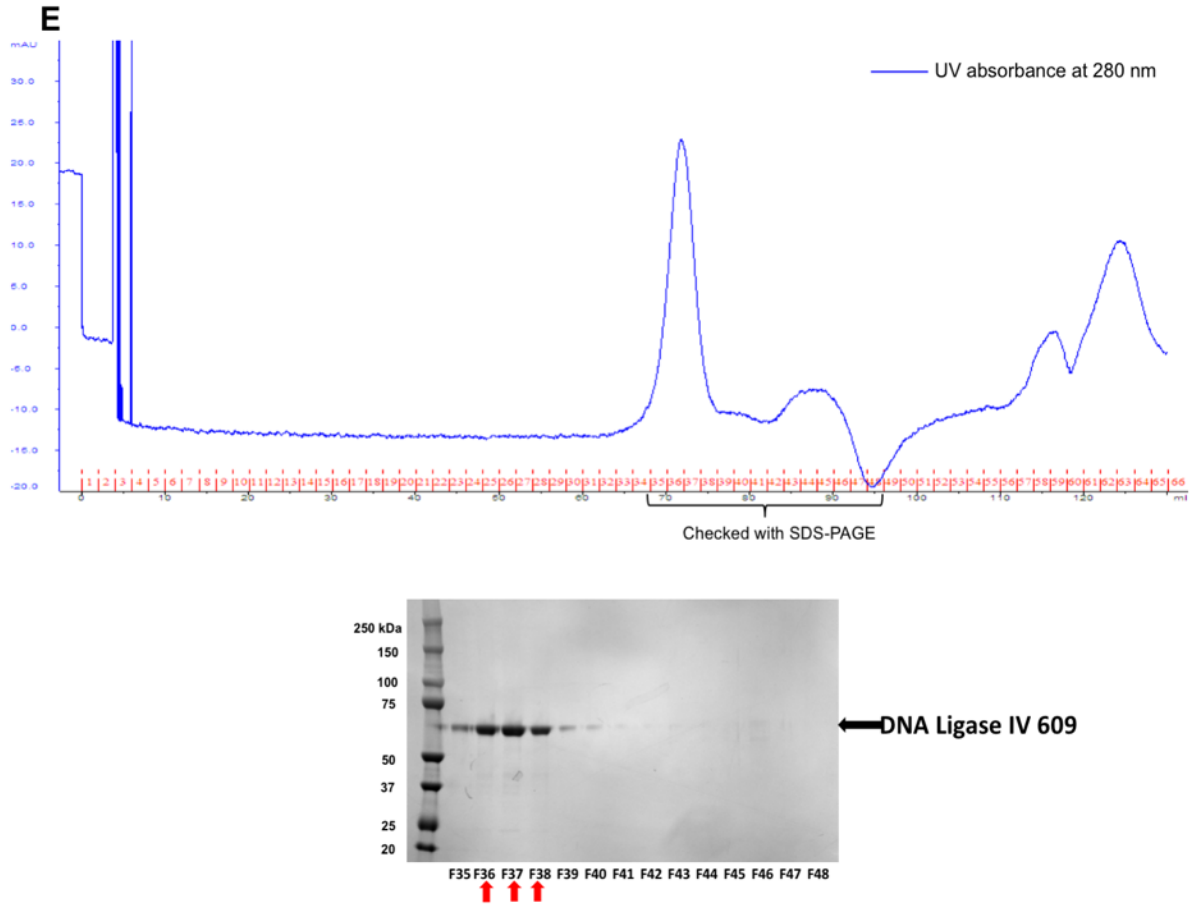
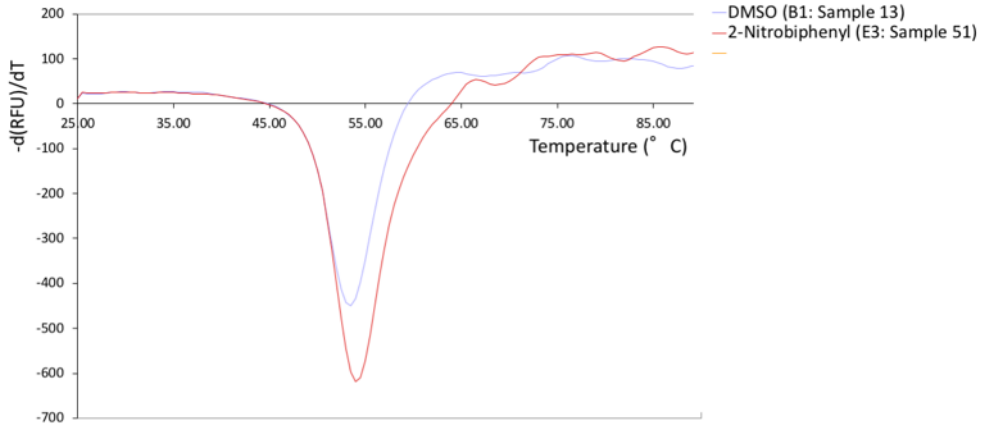
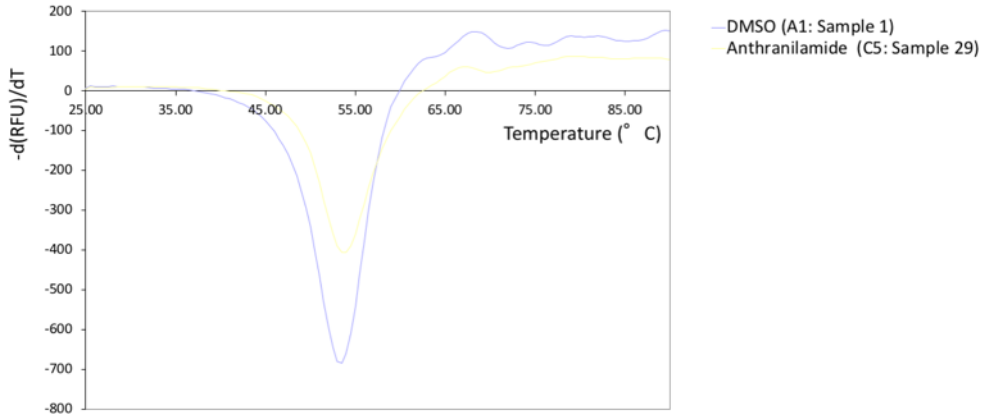
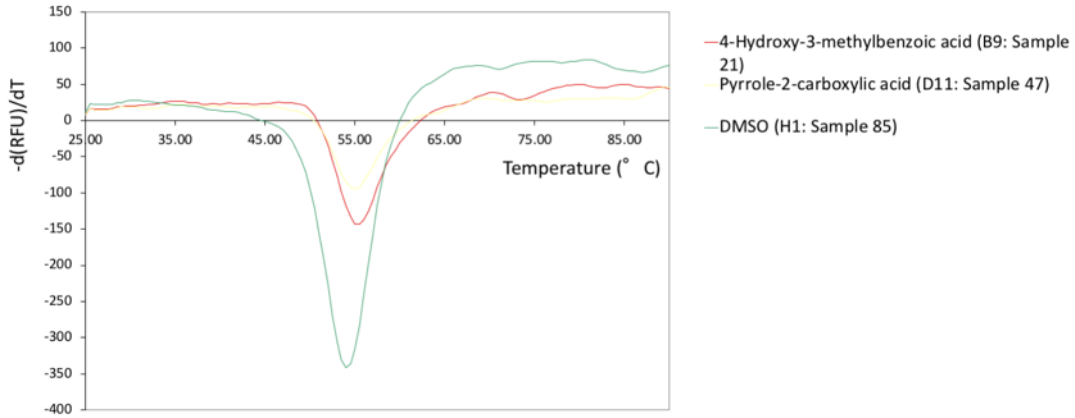
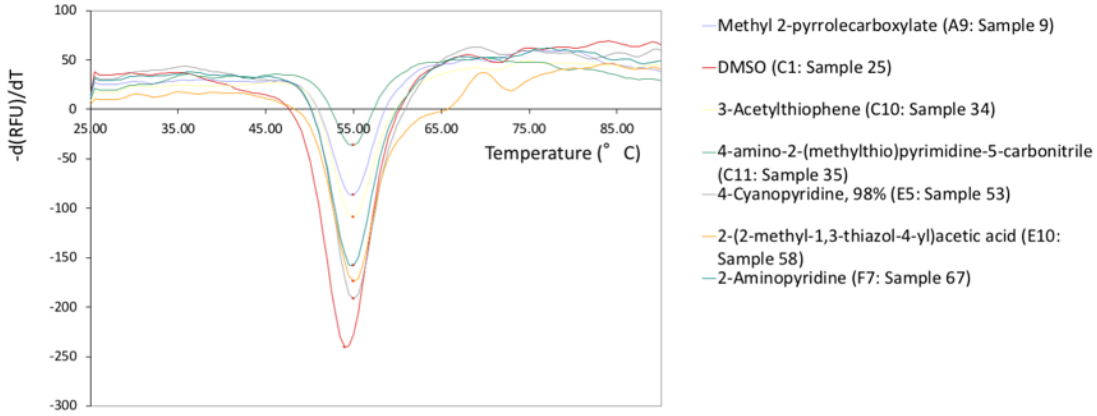


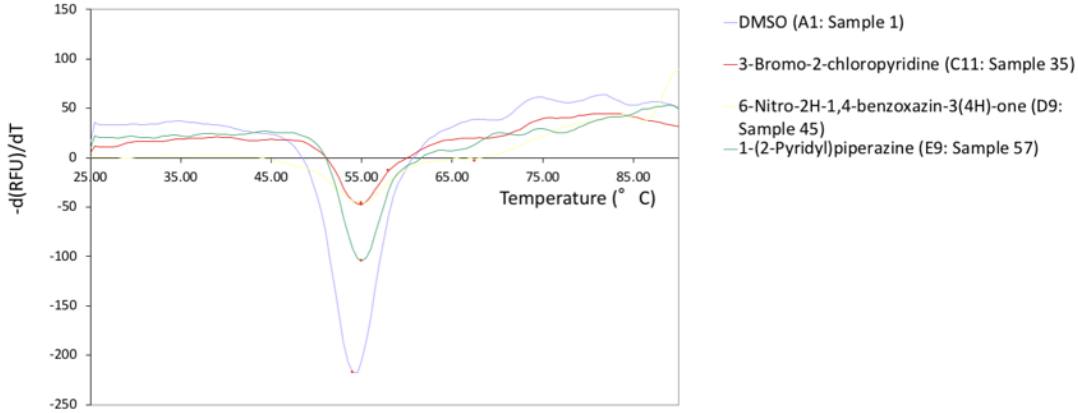
Figure S1. Purification of catalytic domain of DNA ligase IV (DNA Ligase IV 609). (A) His-tag purification of DNA Ligase IV 609 using Ni-NTA; (B) HisTrap-column purification of DNA Ligase IV 609 with gradient elution; (C) Heparin purification of DNA Ligase IV 609 using HiTrap Heparin column; (D) Pheyl purification of DNA Ligase IV 609 using HiTrap Phenyl column. (E) Gel-filtration purification of DNA Ligase IV 609 using Superdex 200 16/60 column. The numbers shown next to the gels are the molecular weights (kDa) of the markers. F stands for Fraction in the lane labels, followed by the fraction number. The red arrow indicates the final collected fraction. All SDS-PAGE gels were stained using Coomassie Blue.

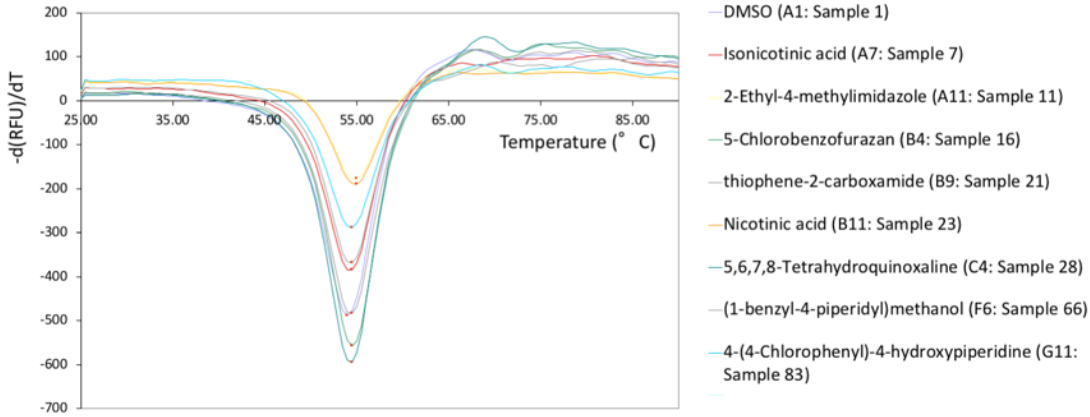

DMSO (B1: Sample 13)
2-Nitrobiphenyl (E3: Sample 51)
-d(RFU)/dT
Temperature (° C)


DMSO (A1: Sample 1)
Anthranilamide (C5: Sample 29)
-d(RFU)/dT
Temperature (° C)


4-Hydroxy-3-methylbenzoic acid (B9: Sample 21)
Pyrrole-2-carboxylic acid (D11: Sample 47)
DMSO (H1: Sample 85)
-d(RFU)/dT
Temperature (° C)


-d(RFU)/dT
Temperature (° C)
Methyl 2-pyrrolecarboxylate (A9: Sample 9)
DMSO (C1: Sample 25)
3-Acetylthiophene (C10: Sample 34)
4-amino-2-(methylthio)pyrimidine-5-carbonitrile (C11: Sample 35)
4-Cyanopyridine, 98% (E5: Sample 53)
2-(2-methyl-1,3-thiazol-4-yl)acetic acid (E10: Sample 58)
2-Aminopyridine (F7: Sample 67)


-d(RFU)/dT
Temperature (° C)
DMSO (A1: Sample 1)
3-Bromo-2-chloropyridine (C11: Sample 35)
6-Nitro-2H-1,4-benzoxazin-3(4H)-one (D9: Sample 45)
1-(2-Pyridyl)piperazine (E9: Sample 57)


-d(RFU)/dT
Temperature (° C)
DMSO (A1: Sample 1)
Isonicotinic acid (A7: Sample 7)
2-Ethyl-4-methylimidazole (A11: Sample 11)
5-Chlorobenzofurazan (B4: Sample 16)
thiophene-2-carboxamide (B9: Sample 21)
Nicotinic acid (B11: Sample 23)
5,6,7,8-Tetrahydroquinoxaline (C4: Sample 28)
(1-benzyl-4-piperidyl)methanol (F6: Sample 66)
4-(4-Chlorophenyl)-4-hydroxypiperidine (G11: Sample 83)

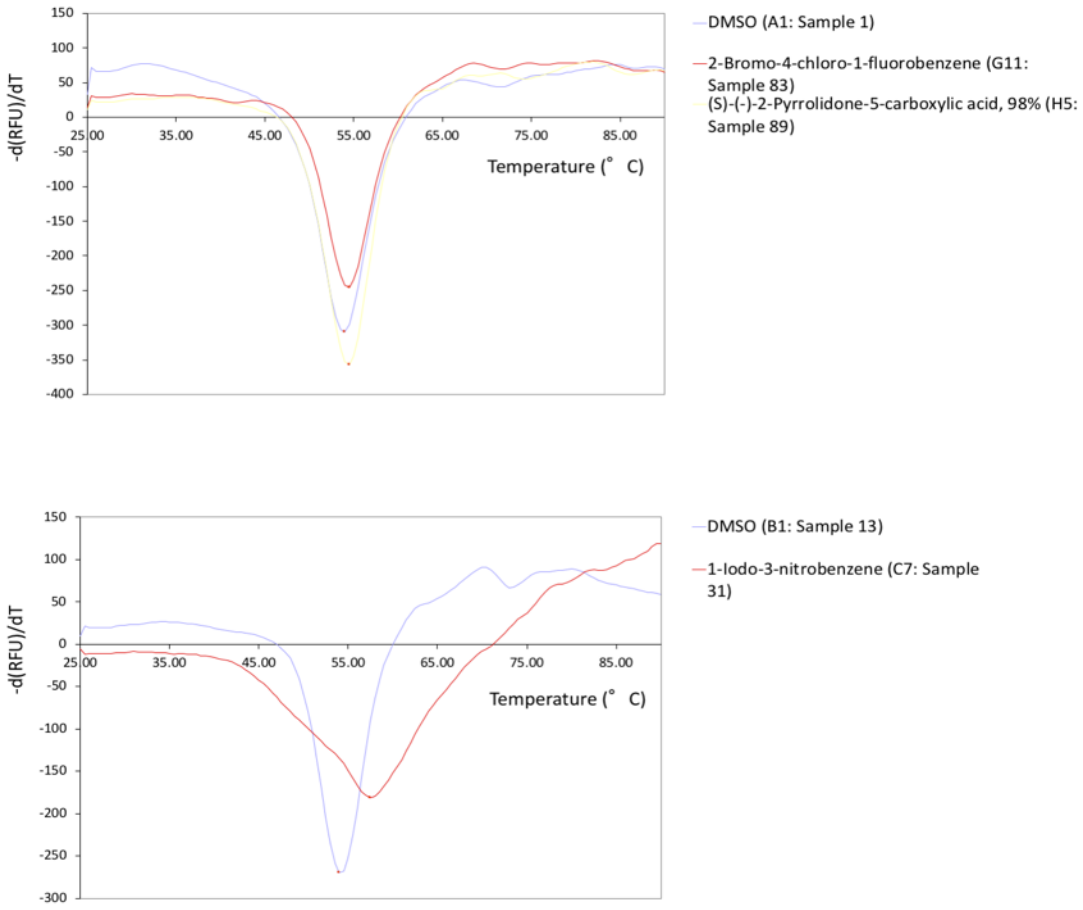


Figure S2. DSF profiles of fragment screening on DNA ligase IV DBD. Only the curves of positive hits and negative control are shown in the profiles.

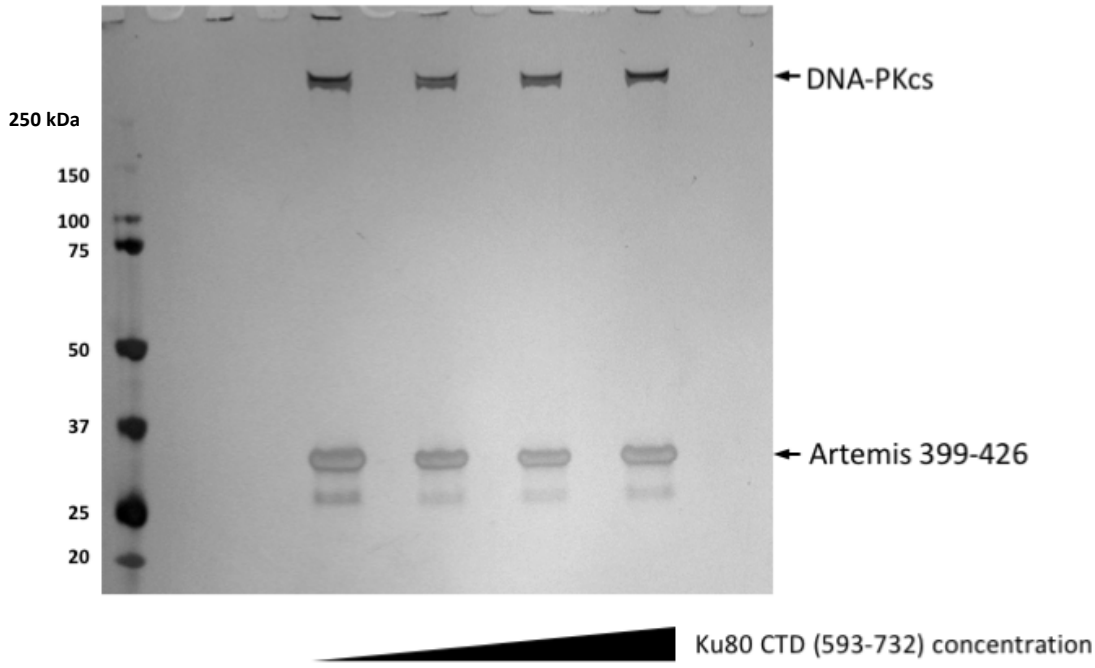


Figure S3. His-tag pulldown assay of DNA-PKcs with Artemis 399-426 and gradient Ku80 CTD (593-732). The triangle labelled with Ku80CTD concentration indicates the gradient of concentration from the low end to the high end. The gels was NuPAGE™ 4-12% Bis-Tris Protein Gels stained with silver staining. The numbers shown next to gel are the molecular weight (kDa) of the markers

References

Ahel, I. *et al.* (2006) 'The neurodegenerative disease protein aprataxin resolves abortive DNA ligation intermediates', *Nature*, 443(7112), pp. 713–716. doi: 10.1038/nature05164.

Ahnesorg, P., Smith, P. and Jackson, S. P. (2006) 'XLF Interacts with the XRCC4-DNA Ligase IV Complex to Promote DNA Nonhomologous End-Joining', *Cell*, 124(2), pp. 301–313. doi: 10.1016/j.cell.2005.12.031.

Alan E. Tomkinson, *,† *et al.* (2006) 'DNA Ligases: Structure, Reaction Mechanism, and Function'. American Chemical Society . doi: 10.1021/CR040498D.

Allegretti, M. *et al.* (2014) 'Atomic model of the F420-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector', *eLife*, 3. doi: 10.7554/eLife.01963.

Allerston, C. K. *et al.* (2015) 'The structures of the SNM1A and SNM1B/Apollo nuclease domains reveal a potential basis for their distinct DNA processing activities.', *Nucleic acids research*. Oxford University Press, 43(22), pp. 11047–60. doi: 10.1093/nar/gkv1256.

Altschul, S. *et al.* (1997) 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Research*, 25(17), pp. 3389–3402. doi: 10.1093/nar/25.17.3389.

Amunts, A. *et al.* (2014) 'Structure of the Yeast Mitochondrial Large Ribosomal Subunit', *Science*, 343(6178), pp. 1485–1489. doi: 10.1126/science.1249410.

An, J. *et al.* (2010) 'DNA-PKcs plays a dominant role in the regulation of H2AX phosphorylation in response to DNA damage and cell cycle progression.', *BMC molecular biology*. BioMed Central, 11, p. 18. doi: 10.1186/1471-2199-11-18.

Andres, S. N. *et al.* (2012) 'A human XRCC4–XLF complex bridges DNA', *Nucleic Acids Research*, 40(4), pp. 1868–1878. doi: 10.1093/nar/gks022.

Arnoult, N. *et al.* (2017) 'Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN', *Nature*. Nature Research. doi: 10.1038/nature24023.

Arosio, D. *et al.* (2002) 'Studies on the Mode of Ku Interaction with DNA', *Journal of Biological Chemistry*, 277(12), pp. 9741–9748. doi: 10.1074/jbc.M111916200.

Balmus, G. *et al.* (2016) 'Synthetic lethality between PAXX and XLF in mammalian development', *Genes & Development*, 30(19), pp. 2152–2157. doi: 10.1101/gad.290510.116.

Bernstein, N. K. *et al.* (2005) 'The Molecular Architecture of the Mammalian DNA Repair Enzyme, Polynucleotide Kinase', *Molecular Cell*, 17(5), pp. 657–670. doi: 10.1016/j.molcel.2005.02.012.

Bétermier, M., Bertrand, P. and Lopez, B. S. (2014) 'Is Non-Homologous End-Joining Really an Inherently Error-Prone Process?', *PLoS Genetics*. Edited by S. Jinks-Robertson. Public Library of Science, 10(1), p. e1004086. doi: 10.1371/journal.pgen.1004086.

Birrane, G. *et al.* (2007) 'Crystal Structure of the BARD1 BRCT Domains † , ‡', *Biochemistry*, 46(26), pp. 7706–7712. doi: 10.1021/bi700323t.

Blackford, A. N. and Jackson, S. P. (2017) 'ATM, ATR, and DNA-PK: The Trinity at the Heart of the DNA Damage Response', *Molecular Cell*, 66(6), pp. 801–817. doi: 10.1016/j.molcel.2017.05.015.

Bork, P. *et al.* (1997) 'A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins.', *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 11(1), pp. 68–76. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9034168 (Accessed: 29 August 2019).

Boskovic, J. *et al.* (2003) 'Visualization of DNA-induced conformational changes in the DNA repair kinase DNA-PKcs.', *The EMBO journal*. European Molecular Biology Organization, 22(21), pp. 5875–82. doi: 10.1093/emboj/cdg555.

Bosotti, R., Isacchi, A. and Sonnhammer, E. L. L. (2000) 'FAT: a novel domain in PIK-related kinases', *Trends in Biochemical Sciences*. Elsevier Current Trends, 25(5), pp. 225–227. doi: 10.1016/S0968-0004(00)01563-2.

Botuyan, M. V. *et al.* (2006) 'Structural Basis for the Methylation State-Specific Recognition of Histone H4-K20 by 53BP1 and Crb2 in DNA Repair', *Cell*, 127(7), pp. 1361–1373. doi: 10.1016/j.cell.2006.10.043.

Buck, D. *et al.* (2006) 'Cernunnos, a Novel Nonhomologous End-Joining Factor, Is Mutated in Human Immunodeficiency with Microcephaly', *Cell*, 124(2), pp. 287–299. doi: 10.1016/j.cell.2005.12.030.

Calsou, P. *et al.* (2003) 'Coordinated assembly of Ku and p460 subunits of the DNA-dependent protein kinase on DNA ends is necessary for XRCC4-ligase IV recruitment.', *Journal of molecular biology*, 326(1), pp. 93–103. doi: 10.1016/s0022-2836(02)01328-1.

Carter, T. *et al.* (1990) 'A DNA-activated protein kinase from HeLa cell nuclei.', *Molecular and cellular biology*, 10(12), pp. 6460–71. doi: 10.1128/mcb.10.12.6460.

Cary, R. B. *et al.* (1997) 'DNA looping by Ku and the DNA-dependent protein kinase', *Proceedings of the National Academy of Sciences*, 94(9), pp. 4267–4272. doi: 10.1073/pnas.94.9.4267.

Ceccaldi, R., Rondinelli, B. and D'Andrea, A. D. (2016) 'Repair Pathway Choices and Consequences at the Double-Strand Break', *Trends in Cell Biology*, 26(1), pp. 52–64. doi: 10.1016/j.tcb.2015.07.009.

Chan, D. W. *et al.* (1999) 'DNA-Dependent Protein Kinase Phosphorylation Sites in Ku 70/80 Heterodimer †', *Biochemistry*, 38(6), pp. 1819–1828. doi: 10.1021/bi982584b.

Chang, H. H. Y. *et al.* (2017) 'Non-homologous DNA end joining and alternative pathways to double-strand break repair', *Nature Reviews Molecular Cell Biology*, 18(8), pp. 495–506. doi: 10.1038/nrm.2017.48.

Chang, H. H. Y. and Lieber, M. R. (2016) 'Structure-Specific nuclease activities of Artemis and the Artemis: DNA-PKcs complex.', *Nucleic acids research*. Oxford University Press, 44(11), pp. 4991–7. doi: 10.1093/nar/gkw456.

Chapman, J. R. *et al.* (2013) 'RIF1 is essential for 53BP1-dependent nonhomologous end joining and suppression of DNA double-strand break resection.', *Molecular cell*. Elsevier, 49(5), pp. 858–71. doi: 10.1016/j.molcel.2013.01.002.

Chen, L. *et al.* (2005) 'ATM and Chk2-dependent phosphorylation of MDMX contribute to p53 activation after DNA damage', *The EMBO Journal*, 24(19), pp. 3411–3422. doi: 10.1038/sj.emboj.7600812.

Chen, S.-H. and Yu, X. (2019) 'Human DNA ligase IV is able to use NAD+ as an alternative adenylation donor for DNA ends ligation', *Nucleic Acids Research*. Narnia, 47(3), pp. 1321–1334. doi: 10.1093/nar/gky1202.

Cherry, A. L. *et al.* (2015) 'Versatility in phospho-dependent molecular recognition of the

XRCC1 and XRCC4 DNA-damage scaffolds by aprataxin-family FHA domains', *DNA Repair*, 35, pp. 116–125. doi: 10.1016/j.dnarep.2015.10.002.

Chirgadze, D. Y. *et al.* (2017) 'DNA-PKcs, Allostery, and DNA Double-Strand Break Repair', in *Methods in enzymology*, pp. 145–157. doi: 10.1016/bs.mie.2017.04.001.

Chiu, C. Y. *et al.* (1998) 'Cryo-EM imaging of the catalytic subunit of the DNA-dependent protein kinase', *Journal of Molecular Biology*, 284(4), pp. 1075–1081. doi: 10.1006/jmbi.1998.2212.

Chou, C. H. *et al.* (1992) 'Role of a major autoepitope in forming the DNA binding site of the p70 (Ku) antigen', *Journal of Experimental Medicine*, 175(6), pp. 1677–1684. doi: 10.1084/jem.175.6.1677.

Clements, P. M. *et al.* (2004) 'The ataxia–oculomotor apraxia 1 gene product has a role distinct from ATM and interacts with the DNA strand break repair proteins XRCC1 and XRCC4', *DNA Repair*, 3(11), pp. 1493–1502. doi: 10.1016/j.dnarep.2004.06.017.

Craxton, A. *et al.* (2018) 'PAXX and its paralogs synergistically direct DNA polymerase λ activity in DNA repair', *Nature Communications*. Nature Publishing Group, 9(1), p. 3877. doi: 10.1038/s41467-018-06127-y.

Cui, X. *et al.* (2005) 'Autophosphorylation of DNA-Dependent Protein Kinase Regulates DNA End Processing and May Also Alter Double-Strand Break Repair Pathway Choice', *Molecular and Cellular Biology*, 25(24), pp. 10842–10852. doi: 10.1128/MCB.25.24.10842-10852.2005.

De Ioannes, P. *et al.* (2012) 'Structural Basis of DNA Ligase IV-Artemis Interaction in Nonhomologous End-Joining', *Cell Reports*, 2(6), pp. 1505–1512. doi: 10.1016/j.celrep.2012.11.004.

De Muyt, A. *et al.* (2012) 'BLM Helicase Ortholog Sgs1 Is a Central Regulator of Meiotic Recombination Intermediate Metabolism', *Molecular Cell*, 46(1), pp. 43–53. doi: 10.1016/j.molcel.2012.02.020.

DeFazio, L. G. *et al.* (2002) 'Synapsis of DNA ends by DNA-dependent protein kinase', *The EMBO Journal*, 21(12), pp. 3192–3200. doi: 10.1093/emboj/cdf299.

Derbyshire, D. J. *et al.* (2002) 'Crystal structure of human 53BP1 BRCT domains bound to p53 tumour suppressor', *The EMBO Journal*, 21(14), pp. 3863–3872. doi: 10.1093/emboj/cdf383.

Dobbs, T. A., Tainer, J. A. and Lees-Miller, S. P. (2010) 'A structural model for regulation of NHEJ by DNA-PKcs autophosphorylation', *DNA Repair*, 9(12), pp. 1307–1314. doi: 10.1016/j.dnarep.2010.09.019.

Dosztanyi, Z. *et al.* (2005) 'IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content', *Bioinformatics*, 21(16), pp. 3433–3434. doi: 10.1093/bioinformatics/bti541.

Dosztányi, Z., Mészáros, B. and Simon, I. (2009) 'ANCHOR: web server for predicting protein binding regions in disordered proteins.', *Bioinformatics (Oxford, England)*. Oxford University Press, 25(20), pp. 2745–6. doi: 10.1093/bioinformatics/btp518.

Douglas, P. *et al.* (2002) 'Identification of in vitro and in vivo phosphorylation sites in the catalytic subunit of the DNA-dependent protein kinase.', *The Biochemical journal*, 368(Pt 1), pp. 243–51. doi: 10.1042/BJ20020973.

Douglas, P. *et al.* (2005) 'DNA-PK-dependent phosphorylation of Ku70/80 is not required for non-homologous end joining', *DNA Repair*, 4(9), pp. 1006–1018. doi: 10.1016/j.dnarep.2005.05.003.

Douglas, P. *et al.* (2007) 'The DNA-Dependent Protein Kinase Catalytic Subunit Is Phosphorylated In Vivo on Threonine 3950, a Highly Conserved Amino Acid in the Protein Kinase Domain', *Molecular and Cellular Biology*, 27(5), pp. 1581–1591. doi: 10.1128/MCB.01962-06.

Drouet, J. *et al.* (2006) 'Interplay between Ku, Artemis, and the DNA-dependent protein kinase catalytic subunit at DNA ends.', *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 281(38), pp. 27784–93. doi: 10.1074/jbc.M603047200.

Drozdetskiy, A. *et al.* (2015) 'JPred4: a protein secondary structure prediction server', *Nucleic Acids Research*, 43(W1), pp. W389–W394. doi: 10.1093/nar/gkv332.

Dunker, A. K. *et al.* (2002) 'Intrinsic disorder and protein function.', *Biochemistry*, 41(21), pp. 6573–82. doi: 10.1021/bi012159+.

Dvir, A. *et al.* (1992) 'Ku autoantigen is the regulatory component of a template-associated protein kinase that phosphorylates RNA polymerase II.', *Proceedings of the National Academy of Sciences*, 89(24), pp. 11920–11924. doi: 10.1073/pnas.89.24.11920.

Dyson, H. J. and Wright, P. E. (2005) 'Intrinsically unstructured proteins and their functions', *Nature Reviews Molecular Cell Biology*, 6(3), pp. 197–208. doi: 10.1038/nrm1589.

Edgar, R. C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*, 32(5), pp. 1792–1797. doi: 10.1093/nar/gkh340.

Ellenberger, T. and Tomkinson, A. E. (2008) 'Eukaryotic DNA Ligases: Structural and Functional Insights', *Annual Review of Biochemistry*, 77(1), pp. 313–338. doi: 10.1146/annurev.biochem.77.061306.123941.

Elmlund, D. and Elmlund, H. (2012) 'SIMPLE: Software for ab initio reconstruction of heterogeneous single-particles', *Journal of Structural Biology*. Academic Press, 180(3), pp. 420–427. doi: 10.1016/J.JSB.2012.07.010.

Emerson, C. H. and Bertuch, A. A. (2016) 'Consider the workhorse: Nonhomologous end-joining in budding yeast.', *Biochemistry and cell biology = Biochimie et biologie cellulaire*. NIH Public Access, 94(5), pp. 396–406. doi: 10.1139/bcb-2016-0001.

Ericsson, U. B. *et al.* (2006) 'Thermofluor-based high-throughput stability optimization of proteins for structural studies', *Analytical Biochemistry*, 357(2), pp. 289–298. doi: 10.1016/j.ab.2006.07.027.

Fan, X. *et al.* (2019) 'Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 Angstrom resolution', *Nature Communications*. Nature Publishing Group, 10(1), p. 2386. doi: 10.1038/s41467-019-10368-w.

Fell, V. L. and Schild-Poulter, C. (2012) 'Ku regulates signaling to DNA damage response pathways through the Ku70 von Willebrand A domain.', *Molecular and cellular biology*, 32(1), pp. 76–87. doi: 10.1128/MCB.05661-11.

Feng, L. and Chen, J. (2012) 'The E3 ligase RNF8 regulates KU80 removal and NHEJ repair', *Nature Structural & Molecular Biology*, 19(2), pp. 201–206. doi: 10.1038/nsmb.2211.

Fradet-Turcotte, A. *et al.* (2013) '53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark', *Nature*, 499(7456), pp. 50–54. doi: 10.1038/nature12318.

Frit, P. *et al.* (2019) 'Plugged into the Ku-DNA hub: The NHEJ network', *Progress in Biophysics and Molecular Biology*. doi: 10.1016/j.pbiomolbio.2019.03.001.

Gell, D. and Jackson, S. P. (1999) 'Mapping of protein-protein interactions within the DNA-

dependent protein kinase complex', *Nucleic Acids Research*, 27(17), pp. 3494–3502. doi: 10.1093/nar/27.17.3494.

Giaccia, A. J. *et al.* (1990) 'Human chromosome 5 complements the DNA double-strand break-repair deficiency and gamma-ray sensitivity of the XR-1 hamster variant.', *American journal of human genetics*, 47(3), pp. 459–69. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1697445 (Accessed: 29 August 2019).

Goodarzi, A. A. *et al.* (2006) 'DNA-PK autophosphorylation facilitates Artemis endonuclease activity', *The EMBO Journal*, 25(16), pp. 3880–3889. doi: 10.1038/sj.emboj.7601255.

Gottlieb, T. M. and Jackson, S. P. (1993) 'The DNA-dependent protein kinase: Requirement for DNA ends and association with Ku antigen', *Cell*. Cell Press, 72(1), pp. 131–142. doi: 10.1016/0092-8674(93)90057-W.

Gouy, M., Guindon, S. and Gascuel, O. (2010) 'SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building', *Molecular Biology and Evolution*, 27(2), pp. 221–224. doi: 10.1093/molbev/msp259.

Grant, T., Rohou, A. and Grigorieff, N. (2018) 'cisTEM, user-friendly software for single-particle image processing', *eLife*, 7. doi: 10.7554/eLife.35383.

Grawunder, U. *et al.* (1998) 'Requirement for an Interaction of XRCC4 with DNA Ligase IV for Wild-type V(D)J Recombination and DNA Double-strand Break Repair *in Vivo*', *Journal of Biological Chemistry*, 273(38), pp. 24708–24714. doi: 10.1074/jbc.273.38.24708.

Greeson, N. T. *et al.* (2008) 'Di-methyl H4 Lysine 20 Targets the Checkpoint Protein Crb2 to Sites of DNA Damage', *Journal of Biological Chemistry*, 283(48), pp. 33168–33174. doi: 10.1074/jbc.M806857200.

Grundy, G. J. *et al.* (2012) 'APLF promotes the assembly and activity of non-homologous end joining protein complexes', *The EMBO Journal*, 32(1), pp. 112–125. doi: 10.1038/emboj.2012.304.

Grundy, G. J. *et al.* (2016) 'The Ku-binding motif is a conserved module for recruitment and stimulation of non-homologous end-joining proteins', *Nature Communications*. Nature Publishing Group, 7(1), p. 11242. doi: 10.1038/ncomms11242.

Guo, Z. *et al.* (2000) 'Requirement for Atr in phosphorylation of Chk1 and cell cycle regulation

in response to DNA replication blocks and UV-damaged DNA in Xenopus egg extracts', *Genes & Development*. Cold Spring Harbor Laboratory Press, 14(21), p. 2745. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC317027/ (Accessed: 28 August 2019).

Guo, Z. *et al.* (2010) 'ATM Activation by Oxidative Stress', *Science*, 330(6003), pp. 517–521. doi: 10.1126/science.1192912.

Hammel, M. *et al.* (2011) 'XRCC4 Protein Interactions with XRCC4-like Factor (XLF) Create an Extended Grooved Scaffold for DNA Ligation and Double Strand Break Repair', *Journal of Biological Chemistry*, 286(37), pp. 32638–32650. doi: 10.1074/jbc.M111.272641.

Hekmat-Nejad, M. *et al.* (no date) 'Xenopus ATR is a replication-dependent chromatin-binding protein required for the DNA replication checkpoint.', *Current biology : CB*, 10(24), pp. 1565–73. doi: 10.1016/s0960-9822(00)00855-1.

Ho, C. K. *et al.* (2010) 'Mus81 and Yen1 Promote Reciprocal Exchange during Mitotic Recombination to Maintain Genome Integrity in Budding Yeast', *Molecular Cell*, 40(6), pp. 988–1000. doi: 10.1016/j.molcel.2010.11.016.

Huang, Y. *et al.* (2009) 'Impact of a hypomorphic Artemis disease allele on lymphocyte development, DNA end processing, and genome stability', *The Journal of Experimental Medicine*, 206(4), pp. 893–908. doi: 10.1084/jem.20082396.

De Ioannes, P. *et al.* (2012) 'Structural basis of DNA ligase IV-Artemis interaction in nonhomologous end-joining.', *Cell reports*, 2(6), pp. 1505–12. doi: 10.1016/j.celrep.2012.11.004.

Isono, M. *et al.* (2017) 'BRCA1 Directs the Repair Pathway to Homologous Recombination by Promoting 53BP1 Dephosphorylation', *Cell Reports*, 18, pp. 520–532. doi: 10.1016/j.celrep.2016.12.042.

Jasin, M. and Rothstein, R. (2013) 'Repair of Strand Breaks by Homologous Recombination', *Cold Spring Harbor Perspectives in Biology*, 5(11), pp. a012740–a012740. doi: 10.1101/cshperspect.a012740.

Jayaram, S. *et al.* (2008) 'Loss of DNA ligase IV prevents recognition of DNA by double-strand break repair proteins XRCC4 and XLF', *Nucleic Acids Research*, 36(18), pp. 5773–5786. doi: 10.1093/nar/gkn552.

Jette, N. and Lees-Miller, S. P. (2015) 'The DNA-dependent protein kinase: A multifunctional protein kinase with roles in DNA double strand break repair and mitosis', *Progress in Biophysics and Molecular Biology*, 117(2–3), pp. 194–205. doi: 10.1016/j.pbiomolbio.2014.12.003.

Jiang, W. *et al.* (2015) 'Differential Phosphorylation of DNA-PKcs Regulates the Interplay between End-Processing and End-Ligation during Nonhomologous End-Joining', *Molecular Cell*, 58(1), pp. 172–185. doi: 10.1016/j.molcel.2015.02.024.

Joo, W. S. (2002) 'Structure of the 53BP1 BRCT region bound to p53 and its comparison to the Brca1 BRCT structure', *Genes & Development*, 16(5), pp. 583–593. doi: 10.1101/gad.959202.

Jubb, H., Blundell, T. L. and Ascher, D. B. (2015) 'Flexibility and small pockets at protein-protein interfaces: New insights into druggability.', *Progress in biophysics and molecular biology*. Elsevier, 119(1), pp. 2–9. doi: 10.1016/j.pbiomolbio.2015.01.009.

Jung, D. and Alt, F. W. (2004) 'Unraveling V(D)J recombination; insights into gene regulation.', *Cell*. Elsevier, 116(2), pp. 299–311. doi: 10.1016/s0092-8674(04)00039-x.

Junop, M. S. *et al.* (2000) 'Crystal structure of the Xrcc4 DNA repair protein and implications for end joining.', *The EMBO journal*. European Molecular Biology Organization, 19(22), pp. 5962–70. doi: 10.1093/emboj/19.22.5962.

Kaminski, A. M. *et al.* (2018) 'Structures of DNA-bound human ligase IV catalytic core reveal insights into substrate binding and catalysis', *Nature Communications*. Nature Publishing Group, 9(1), p. 2642. doi: 10.1038/s41467-018-05024-8.

Kawale, A. S. and Povirk, L. F. (2018) 'Tyrosyl-DNA phosphodiesterases: rescuing the genome from the risks of relaxation.', *Nucleic acids research*. Oxford University Press, 46(2), pp. 520–537. doi: 10.1093/nar/gkx1219.

Keeney, S. (2008) 'Spo11 and the Formation of DNA Double-Strand Breaks in Meiosis.', *Genome dynamics and stability*. NIH Public Access, 2, pp. 81–123. doi: 10.1007/7050_2007_026.

Kim, J.-S. *et al.* (2005) 'Independent and sequential recruitment of NHEJ and HR factors to DNA damage sites in mammalian cells', *The Journal of Cell Biology*, 170(3), pp. 341–347. doi: 10.1083/jcb.200411083.

Klein, H. L. and Symington, L. S. (2012) 'Sgs1—The Maestro of Recombination', *Cell*, 149(2), pp. 257–259. doi: 10.1016/j.cell.2012.03.020.

Koch, C. A. *et al.* (2004) 'Xrcc4 physically links DNA end processing by polynucleotide kinase to DNA ligation by DNA ligase IV', *The EMBO Journal*, 23(19), pp. 3874–3885. doi: 10.1038/sj.emboj.7600375.

Kumar, V., Alt, F. W. and Frock, R. L. (2016) 'PAXX and XLF DNA repair factors are functionally redundant in joining DNA breaks in a G1-arrested progenitor B-cell line', *Proceedings of the National Academy of Sciences*, 113(38), pp. 10619–10624. doi: 10.1073/pnas.1611882113.

de la Rosa-Trevín, J. M. *et al.* (2016) 'Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy', *Journal of Structural Biology*, 195(1), pp. 93–99. doi: 10.1016/j.jsb.2016.04.010.

Lam, I. and Keeney, S. (2015) 'Mechanism and Regulation of Meiotic Recombination Initiation', *Cold Spring Harbor Perspectives in Biology*, 7(1), p. a016634. doi: 10.1101/cshperspect.a016634.

Langerak, P. *et al.* (2011) 'Release of Ku and MRN from DNA Ends by Mre11 Nuclease Activity and Ctp1 Is Required for Homologous Recombination Repair of Double-Strand Breaks', *PLoS Genetics*. Edited by G. P. Copenhaver, 7(9), p. e1002271. doi: 10.1371/journal.pgen.1002271.

Lee, J.-H. and Paull, T. T. (2004) 'Direct Activation of the ATM Protein Kinase by the Mre11/Rad50/Nbs1 Complex', *Science*, 304(5667), pp. 93–96. doi: 10.1126/science.1091496.

Lee, J.-H. and Paull, T. T. (2005) 'ATM Activation by DNA Double-Strand Breaks Through the Mre11-Rad50-Nbs1 Complex', *Science*, 308(5721), pp. 551–554. doi: 10.1126/science.1108297.

Lee, K.-J. *et al.* (2004) 'Identification of DNA-PKcs phosphorylation sites in XRCC4 and effects of mutations at these sites on DNA end joining in a cell-free system', *DNA Repair*, 3(3), pp. 267–276. doi: 10.1016/j.dnarep.2003.11.005.

Lescale, C. *et al.* (2016) 'Specific Roles of XRCC4 Paralogs PAXX and XLF during V(D)J Recombination.', *Cell reports*. Elsevier, 16(11), pp. 2967–2979. doi: 10.1016/j.celrep.2016.08.069.

Leuther, K. K. *et al.* (1999) 'Structure of DNA-dependent protein kinase: implications for its

regulation by DNA', *The EMBO Journal*, 18(5), pp. 1114–1123. doi: 10.1093/emboj/18.5.1114.

Li, B. and Comai, L. (2000) 'Functional Interaction between Ku and the Werner Syndrome Protein in DNA End Processing', *Journal of Biological Chemistry*, 275(37), pp. 28349–28352. doi: 10.1074/jbc.C000289200.

Li, B. and Comai, L. (2001) 'Requirements for the Nucleolytic Processing of DNA Ends by the Werner Syndrome Protein-Ku70/80 Complex', *Journal of Biological Chemistry*, 276(13), pp. 9896–9902. doi: 10.1074/jbc.M008575200.

Li, S. *et al.* (2014) 'Evidence That the DNA Endonuclease ARTEMIS also Has Intrinsic 5'-Exonuclease Activity', *Journal of Biological Chemistry*, 289(11), pp. 7825–7834. doi: 10.1074/jbc.M113.544874.

Li, X. *et al.* (2013) 'Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM', *Nature Methods*. Nature Publishing Group, 10(6), pp. 584–590. doi: 10.1038/nmeth.2472.

Li, Z. *et al.* (1995) 'The XRCC4 gene encodes a novel protein involved in DNA double-strand break repair and V(D)J recombination', *Cell*, 83(7), pp. 1079–1089. doi: 10.1016/0092-8674(95)90135-3.

Liang, S. *et al.* (2016). Achieving selectivity in space and time with DNA double-strand-break response and repair: molecular stages and scaffolds come with strings attached. Structural Chemistry, 28, 161-171. doi: 10.1007/s11224-016-0841-7.

Liao, M. *et al.* (2013) 'Structure of the TRPV1 ion channel determined by electron cryo-microscopy', *Nature*, 504(7478), pp. 107–112. doi: 10.1038/nature12822.

Lieber, M. R. (2010) 'The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway', *Annual Review of Biochemistry*, 79(1), pp. 181–211. doi: 10.1146/annurev.biochem.052308.093131.

Linding, R. *et al.* (2003) 'Protein disorder prediction: implications for structural proteomics.', *Structure (London, England : 1993)*, 11(11), pp. 1453–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14604535 (Accessed: 29 August 2019).

Liu, L. *et al.* (2009) 'SNM1B/Apollo interacts with Astrin and is required for the prophase cell cycle checkpoint', *Cell Cycle*, 8(4), pp. 628–638. doi: 10.4161/cc.8.4.7791.

Liu, Q. *et al.* (2000) 'Chk1 is an essential kinase that is regulated by Atr and required for the G(2)/M DNA damage checkpoint.', *Genes & development*, 14(12), pp. 1448–59. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10859164 (Accessed: 28 August 2019).

Liu, X. *et al.* (2017) 'PAXX promotes KU accumulation at DNA breaks and is essential for end-joining in XLF-deficient mice', *Nature Communications*. Nature Publishing Group, 8(1), p. 13816. doi: 10.1038/ncomms13816.

Lodish, H. F. (2000) *Molecular cell biology*. W.H. Freeman.

Loeffler, P. A. *et al.* (2011) 'Structural studies of the PARP-1 BRCT domain', *BMC Structural Biology*, 11(1), p. 37. doi: 10.1186/1472-6807-11-37.

Ma, Y. *et al.* (2002) 'Hairpin Opening and Overhang Processing by an Artemis/DNA-Dependent Protein Kinase Complex in Nonhomologous End Joining and V(D)J Recombination', *Cell*, 108(6), pp. 781–794. doi: 10.1016/S0092-8674(02)00671-2.

Ma, Y. *et al.* (2004) 'A Biochemically Defined System for Mammalian Nonhomologous DNA End Joining', *Molecular Cell*, 16(5), pp. 701–713. doi: 10.1016/j.molcel.2004.11.017.

Ma, Y. *et al.* (2005) 'The DNA-dependent protein kinase catalytic subunit phosphorylation sites in human Artemis.', *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 280(40), pp. 33839–46. doi: 10.1074/jbc.M507113200.

Macrae, C. J. *et al.* (2008) 'APLF (C2orf13) facilitates nonhomologous end-joining and undergoes ATM-dependent hyperphosphorylation following ionizing radiation', *DNA Repair*, 7(2), pp. 292–302. doi: 10.1016/j.dnarep.2007.10.008.

Mahaney, B. L. *et al.* (2013) 'XRCC4 and XLF form long helical protein filaments suitable for DNA end protection and alignment to facilitate DNA double strand break repair.', *Biochemistry and cell biology = Biochimie et biologie cellulaire*. NIH Public Access, 91(1), pp. 31–41. doi: 10.1139/bcb-2012-0058.

Malu, S. *et al.* (2012) 'Artemis C-terminal region facilitates V(D)J recombination through its interactions with DNA Ligase IV and DNA-PKcs', *The Journal of Experimental Medicine*, 209(5), pp. 955–963. doi: 10.1084/jem.20111437.

Manke, I. A. (2003) 'BRCT Repeats As Phosphopeptide-Binding Modules Involved in Protein Targeting', *Science*, 302(5645), pp. 636–639. doi: 10.1126/science.1088877.

Maréchal, A. and Zou, L. (2013) 'DNA damage sensing by the ATM and ATR kinases.', *Cold Spring Harbor perspectives in biology*. Cold Spring Harbor Laboratory Press, 5(9). doi: 10.1101/cshperspect.a012716.

Mari, P.-O. *et al.* (2006) 'Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4', *Proceedings of the National Academy of Sciences*, 103(49), pp. 18597–18602. doi: 10.1073/pnas.0609061103.

Matos, J. *et al.* (2011) 'Regulatory Control of the Resolution of DNA Recombination Intermediates during Meiosis and Mitosis', *Cell*, 147(1), pp. 158–172. doi: 10.1016/j.cell.2011.08.032.

Matsuoka, S. *et al.* (2007) 'ATM and ATR Substrate Analysis Reveals Extensive Protein Networks Responsive to DNA Damage', *Science*, 316(5828), pp. 1160–1166. doi: 10.1126/science.1140321.

Matthews, L. A. and Simmons, L. A. (2014) 'Bacterial nonhomologous end joining requires teamwork.', *Journal of bacteriology*. American Society for Microbiology Journals, 196(19), pp. 3363–5. doi: 10.1128/JB.02042-14.

Meek, K. *et al.* (2007) 'trans Autophosphorylation at DNA-dependent protein kinase's two major autophosphorylation site clusters facilitates end processing but not end joining.', *Molecular and cellular biology*. American Society for Microbiology Journals, 27(10), pp. 3881–90. doi: 10.1128/MCB.02366-06.

Menon, V. and Povirk, L. F. (2017) 'XLF/Cernunnos: An important but puzzling participant in the nonhomologous end joining DNA repair pathway', *DNA Repair*, 58, pp. 29–37. doi: 10.1016/j.dnarep.2017.08.003.

Milligan, J. R. *et al.* (1995) 'DNA Repair by Thiols in Air Shows Two Radicals Make a Double-Strand Break', *Radiation Research*. Radiation Research Society , 143(3), p. 273. doi: 10.2307/3579213.

Mizuta, R. *et al.* (1997) 'Molecular genetic characterization of XRCC4 function', *International Immunology*, 9(10), pp. 1607–1613. doi: 10.1093/intimm/9.10.1607.

Modesti, M. *et al.* (2003) 'Tetramerization and DNA Ligase IV Interaction of the DNA Double-strand Break Repair Protein XRCC4 are Mutually Exclusive', *Journal of Molecular Biology*,

334(2), pp. 215–228. doi: 10.1016/j.jmb.2003.09.031.

Modesti, M., Hesse, J. E. and Gellert, M. (1999) 'DNA binding of Xrcc4 protein is associated with V(D)Jrecombination but not with stimulation of DNA ligase IV activity', *The EMBO Journal*, 18(7), pp. 2008–2018. doi: 10.1093/emboj/18.7.2008.

Mordes, D. A. *et al.* (2008) 'TopBP1 activates ATR through ATRIP and a PIKK regulatory domain.', *Genes & development*. Cold Spring Harbor Laboratory Press, 22(11), pp. 1478–89. doi: 10.1101/gad.1666208.

Moriya, T. *et al.* (2017) 'High-resolution Single Particle Analysis from Electron Cryo-microscopy Images Using SPHIRE', *Journal of Visualized Experiments*, (123), p. e55448. doi: 10.3791/55448.

Moscariello, M. *et al.* (2015) 'Role for Artemis nuclease in the repair of radiation-induced DNA double strand breaks by alternative end joining', *DNA Repair*, 31, pp. 29–40. doi: 10.1016/j.dnarep.2015.04.004.

Moshous, D. *et al.* (2001) 'Artemis, a novel DNA double-strand break repair/V(D)J recombination protein, is mutated in human severe combined immune deficiency.', *Cell*, 105(2), pp. 177–86. doi: 10.1016/s0092-8674(01)00309-9.

Nakamura, K. *et al.* (2019) 'H4K20me0 recognition by BRCA1–BARD1 directs homologous recombination to sister chromatids', *Nature Cell Biology*. Nature Publishing Group, 21(3), pp. 311–318. doi: 10.1038/s41556-019-0282-9.

Neale, M. J., Pan, J. and Keeney, S. (2005) 'Endonucleolytic processing of covalent protein-linked DNA double-strand breaks', *Nature*. Nature Publishing Group, 436(7053), pp. 1053–1057. doi: 10.1038/nature03872.

Nemoz, C. *et al.* (2018a) 'XLF and APLF bind Ku80 at two remote sites to ensure DNA repair by non-homologous end joining', *Nature Structural & Molecular Biology*, 25(10), pp. 971–980. doi: 10.1038/s41594-018-0133-6.

Nemoz, C. *et al.* (2018b) 'XLF and APLF bind Ku80 at two remote sites to ensure DNA repair by non-homologous end joining', *Nature Structural & Molecular Biology*, 25(10), pp. 971–980. doi: 10.1038/s41594-018-0133-6.

Niewolik, D. *et al.* (2017) 'Autoinhibition of the Nuclease ARTEMIS Is Mediated by a Physical

Interaction between Its Catalytic and C-terminal Domains', *Journal of Biological Chemistry*, 292(8), pp. 3351–3365. doi: 10.1074/jbc.M116.770461.

Nimonkar, A. V. *et al.* (2011) 'BLM–DNA2–RPA–MRN and EXO1–BLM–RPA–MRN constitute two DNA end resection machineries for human DNA break repair', *Genes & Development*. Cold Spring Harbor Laboratory Press, 25(4), p. 350. doi: 10.1101/GAD.2003811.

O'Connor, M. J. (2015) 'Targeting the DNA Damage Response in Cancer', *Molecular Cell*, 60(4), pp. 547–560. doi: 10.1016/j.molcel.2015.10.040.

Ochi, T. *et al.* (2015) 'PAXX, a paralog of XRCC4 and XLF, interacts with Ku to promote DNA double-strand break repair', *Science*, 347(6218), pp. 185–188. doi: 10.1126/science.1261971.

Ochi, T., Gu, X. and Blundell, T. L. (2013) 'Structure of the Catalytic Region of DNA Ligase IV in Complex with an Artemis Fragment Sheds Light on Double-Strand Break Repair', *Structure/Folding and Design*, 21, pp. 672–679. doi: 10.1016/j.str.2013.02.014.

Ochi, T., Wu, Q. and Blundell, T. L. (2014) 'The spatial organization of non-homologous end joining: from bridging to end joining.', *DNA repair*. Elsevier, 17(100), pp. 98–109. doi: 10.1016/j.dnarep.2014.02.010.

Olcina, M. M. *et al.* (2013) 'Replication Stress and Chromatin Context Link ATM Activation to a Role in DNA Replication', *Molecular Cell*, 52(5), pp. 758–766. doi: 10.1016/j.molcel.2013.10.019.

Ono, A. *et al.* (1994) 'Subtype Analysis of HTLV-1 in Patients with HTLV-1 Uveitis', *Japanese Journal of Cancer Research*. John Wiley & Sons, Ltd (10.1111), 85(8), pp. 767–770. doi: 10.1111/j.1349-7006.1994.tb02945.x.

Pang, D. *et al.* (1997) 'Ku proteins join DNA fragments as shown by atomic force microscopy.', *Cancer research*, 57(8), pp. 1412–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9108436 (Accessed: 28 August 2019).

Pannicke, U. *et al.* (2004) 'Functional and biochemical dissection of the structure-specific nuclease ARTEMIS', *The EMBO Journal*, 23(9), pp. 1987–1997. doi: 10.1038/sj.emboj.7600206.

Pantoliano, M. W. *et al.* (2001) 'High-Density Miniaturized Thermal Shift Assays as a General Strategy for Drug Discovery', *Journal of Biomolecular Screening*, 6(6), pp. 429–440. doi: 10.1177/108705710100600609.

Pascal, J. M. *et al.* (2004) 'Human DNA ligase I completely encircles and partially unwinds nicked DNA', *Nature*. Nature Publishing Group, 432(7016), pp. 473–478. doi: 10.1038/nature03082.

Perry, J. and Kleckner, N. (2003) 'The ATRs, ATMs, and TORs Are Giant HEAT Repeat Proteins', *Cell*. Cell Press, 112(2), pp. 151–155. doi: 10.1016/S0092-8674(03)00033-3.

Pitcher, R. S., Brissett, N. C. and Doherty, A. J. (2007) 'Nonhomologous End-Joining in Bacteria: A Microbial Perspective', *Annual Review of Microbiology*, 61(1), pp. 259–282. doi: 10.1146/annurev.micro.61.080706.093354.

Postow, L. *et al.* (2008) 'Ku80 removal from DNA through double strand break–induced ubiquitylation', *The Journal of Cell Biology*, 182(3), pp. 467–479. doi: 10.1083/jcb.200802146.

Prilusky, J. *et al.* (2005) 'FoldIndex(C): a simple tool to predict whether a given protein sequence is intrinsically unfolded', *Bioinformatics*. Narnia, 21(16), pp. 3435–3438. doi: 10.1093/bioinformatics/bti537.

Punjani, A. *et al.* (2017) 'cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination', *Nature Methods*, 14(3), pp. 290–296. doi: 10.1038/nmeth.4169.

Radoux, C. J. *et al.* (2016) 'Identifying Interactions that Determine Fragment Binding at Protein Hotspots', *Journal of Medicinal Chemistry*, 59(9), pp. 4314–4325. doi: 10.1021/acs.jmedchem.5b01980.

Rappas, M., Oliver, A. W. and Pearl, L. H. (2011) 'Structure and function of the Rad9-binding region of the DNA-damage checkpoint adaptor TopBP1', *Nucleic Acids Research*, 39(1), pp. 313–324. doi: 10.1093/nar/gkq743.

Reid, D. A. *et al.* (2015) 'Organization and dynamics of the nonhomologous end-joining machinery during DNA double-strand break repair'. doi: 10.1073/pnas.1420115112.

Rivera-Calzada, A. *et al.* (2005) 'Three-Dimensional Structure and Regulation of the DNA-Dependent Protein Kinase Catalytic Subunit (DNA-PKcs)', *Structure*, 13(2), pp. 243–255. doi: 10.1016/j.str.2004.12.006.

Rivera-Calzada, A. *et al.* (2007) 'Structural model of full-length human Ku70-Ku80 heterodimer and its recognition of DNA and DNA-PKcs.', *EMBO reports*. European Molecular Biology Organization, 8(1), pp. 56–62. doi: 10.1038/sj.embor.7400847.

Rodriguez, M. *et al.* (2003) 'Phosphopeptide Binding Specificities of BRCA1 COOH-terminal (BRCT) Domains', *Journal of Biological Chemistry*, 278(52), pp. 52914–52918. doi: 10.1074/jbc.C300407200.

Ropars, V. *et al.* (2011) 'Structural characterization of filaments formed by human Xrcc4-Cernunnos/XLF complex involved in nonhomologous DNA end-joining', *Proceedings of the National Academy of Sciences*, 108(31), pp. 12663–12668. doi: 10.1073/pnas.1100758108.

Roth, D. B. (2014) 'V(D)J Recombination: Mechanism, Errors, and Fidelity.', *Microbiology spectrum*. NIH Public Access, 2(6). doi: 10.1128/microbiolspec.MDNA3-0041-2014.

Roy, S. *et al.* (2012a) 'XRCC4's interaction with XLF is required for coding (but not signal) end joining', *Nucleic Acids Research*. Oxford University Press, 40(4), pp. 1684–1694. doi: 10.1093/nar/gkr1315.

Roy, S. *et al.* (2012b) 'XRCC4's interaction with XLF is required for coding (but not signal) end joining', *Nucleic Acids Research*, 40(4), pp. 1684–1694. doi: 10.1093/nar/gkr1315.

Rulten, S. L. and Grundy, G. J. (2017) 'Non-homologous end joining: Common interaction sites and exchange of multiple factors in the DNA repair process', *BioEssays*, 39(3), p. 1600209. doi: 10.1002/bies.201600209.

Šali, A. and Blundell, T. L. (1993) 'Comparative Protein Modelling by Satisfaction of Spatial Restraints', *Journal of Molecular Biology*, 234(3), pp. 779–815. doi: 10.1006/jmbi.1993.1626.

Sartori, A. A. *et al.* (2007) 'Human CtIP promotes DNA end resection.', *Nature*. Europe PMC Funders, 450(7169), pp. 509–14. doi: 10.1038/nature06337.

Scheres, S. H. W. (2012) 'RELION: Implementation of a Bayesian approach to cryo-EM structure determination', *Journal of Structural Biology*. Academic Press, 180(3), pp. 519–530. doi: 10.1016/J.JSB.2012.09.006.

Scheres, S. H. W. (2016) 'Processing of Structurally Heterogeneous Cryo-EM Data in RELION', *Methods in Enzymology*. Academic Press, 579, pp. 125–157. doi: 10.1016/BS.MIE.2016.04.012.

Semisotnov, G. V. *et al.* (1991) 'Study of the ?molten globule? intermediate state in protein folding by a hydrophobic fluorescent probe', *Biopolymers*, 31(1), pp. 119–128. doi: 10.1002/bip.360310111.

Sengerová, B. *et al.* (2012) 'Characterization of the Human SNM1A and SNM1B/Apollo DNA Repair Exonucleases', *Journal of Biological Chemistry*, 287(31), pp. 26254–26267. doi: 10.1074/jbc.M112.367243.

Shamanna, R. A. *et al.* (2016) 'WRN regulates pathway choice between classical and alternative non-homologous end joining', *Nature Communications*. Nature Publishing Group, 7(1), p. 13785. doi: 10.1038/ncomms13785.

Shao, N. *et al.* (2013) 'An updated meta-analysis of XRCC4 polymorphisms and cancer risk based on 31 case-control studies', *Cancer Biomarkers*, 12(1), pp. 37–47. doi: 10.3233/CBM-120292.

Sharif, H. *et al.* (2017) 'Cryo-EM structure of the DNA-PK holoenzyme.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(28), pp. 7367–7372. doi: 10.1073/pnas.1707386114.

Shi, J., Blundell, T. L. and Mizuguchi, K. (2001) 'FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties11Edited by B. Honig', *Journal of Molecular Biology*, 310(1), pp. 243–257. doi: 10.1006/jmbi.2001.4762.

Shibata, A. *et al.* (2011) 'Factors determining DNA double-strand break repair pathway choice in G2 phase', *The EMBO Journal*, 30(6), pp. 1079–1092. doi: 10.1038/emboj.2011.27.

Shuman, S. (2009) 'DNA ligases: progress and prospects.', *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 284(26), pp. 17365–9. doi: 10.1074/jbc.R900017200.

Shuman, S. and Lima, C. D. (2004) 'The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases', *Current Opinion in Structural Biology*, 14(6), pp. 757–764. doi: 10.1016/j.sbi.2004.10.006.

Sibanda, B. L. *et al.* (2001) 'Crystal structure of an Xrcc4-DNA ligase IV complex.', *Nature Structural Biology*, 8(12), pp. 1015–1019. doi: 10.1038/nsb725.

Sibanda, B. L. *et al.* (2017) 'DNA-PKcs structure suggests an allosteric mechanism modulating DNA double-strand break repair', *Science*, 355(6324), pp. 520–524. doi: 10.1126/science.aak9654.

Sibanda, B. L., Chirgadze, D. Y. and Blundell, T. L. (2010) 'Crystal structure of DNA-PKcs reveals a large open-ring cradle comprised of HEAT repeats', *Nature*, 463(7277), pp. 118–121. doi: 10.1038/nature08648.

Singleton, B. K. *et al.* (1997) 'Molecular and biochemical characterization of xrs mutants defective in Ku80.', *Molecular and cellular biology*, 17(3), pp. 1264–73. doi: 10.1128/mcb.17.3.1264.

Soding, J., Biegert, A. and Lupas, A. N. (2005) 'The HHpred interactive server for protein homology detection and structure prediction', *Nucleic Acids Research*. Narnia, 33(Web Server), pp. W244–W248. doi: 10.1093/nar/gki408.

Soubeyrand, S. *et al.* (2006) 'Artemis Phosphorylated by DNA-dependent Protein Kinase Associates Preferentially with Discrete Regions of Chromatin', *Journal of Molecular Biology*, 358(5), pp. 1200–1211. doi: 10.1016/j.jmb.2006.02.061.

Spagnolo, L. *et al.* (2006) 'Three-dimensional structure of the human DNA-PKcs/Ku70/Ku80 complex assembled on DNA and its implications for DNA DSB repair.', *Molecular cell*. Elsevier, 22(4), pp. 511–9. doi: 10.1016/j.molcel.2006.04.013.

Srivastava, M. and Raghavan, S. C. (2015) 'DNA double-strand break repair inhibitors as cancer therapeutics.', *Chemistry & biology*. Elsevier, 22(1), pp. 17–29. doi: 10.1016/j.chembiol.2014.11.013.

Suberbielle, E. *et al.* (2013) 'Physiologic brain activity causes DNA double-strand breaks in neurons, with exacerbation by amyloid-β', *Nature Neuroscience*, 16(5), pp. 613–621. doi: 10.1038/nn.3356.

Sy, S. M. H., Huen, M. S. Y. and Chen, J. (2009) 'PALB2 is an integral component of the BRCA complex required for homologous recombination repair', *Proceedings of the National Academy of Sciences*, 106(17), pp. 7155–7160. doi: 10.1073/pnas.0811159106.

Syeda, A. H., Hawkins, M. and McGlynn, P. (2014) 'Recombination and Replication', *Cold Spring Harbor Perspectives in Biology*, 6(11), pp. a016550–a016550. doi: 10.1101/cshperspect.a016550.

Tang, G. *et al.* (2007) 'EMAN2: An extensible image processing suite for electron microscopy', *Journal of Structural Biology*, 157(1), pp. 38–46. doi: 10.1016/j.jsb.2006.05.009.

Thomas, S. E. *et al.* (2019) 'Structure-guided fragment-based drug discovery at the synchrotron: screening binding sites and correlations with hotspot mapping', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. The Royal Society Publishing, 377(2147), p. 20180422. doi: 10.1098/rsta.2018.0422.

Thompson, L. H. (2012) 'Recognition, signaling, and repair of DNA double-strand breaks produced by ionizing radiation in mammalian cells: The molecular choreography', *Mutation Research/Reviews in Mutation Research*. Elsevier, 751(2), pp. 158–246. doi: 10.1016/J.MRREV.2012.06.002.

Turchi, J. J. and Henkels, K. (1996) 'Human Ku Autoantigen Binds Cisplatin-damaged DNA but Fails to Stimulate Human DNA-activated Protein Kinase', *Journal of Biological Chemistry*, 271(23), pp. 13861–13867. doi: 10.1074/jbc.271.23.13861.

Uematsu, N. *et al.* (2007) 'Autophosphorylation of DNA-PK $_{CS}$ regulates its dynamics at DNA double-strand breaks', *The Journal of Cell Biology*, 177(2), pp. 219–229. doi: 10.1083/jcb.200608077.

de Villartay, J.-P. *et al.* (2009) 'A histidine in the β-CASP domain of Artemis is critical for its full in vitro and in vivo functions', *DNA Repair*, 8(2), pp. 202–208. doi: 10.1016/j.dnarep.2008.10.010.

Walker, J. R., Corpina, R. A. and Goldberg, J. (2001) 'Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair', *Nature*, 412(6847), pp. 607–614. doi: 10.1038/35088000.

Wang, J. *et al.* (2014) 'PTIP associates with Artemis to dictate DNA repair pathway choice.', *Genes & development*. Cold Spring Harbor Laboratory Press, 28(24), pp. 2693–8. doi: 10.1101/gad.252478.114.

Wang, J., Dong, X. and Reeves, W. H. (1998) 'A model for Ku heterodimer assembly and interaction with DNA. Implications for the function of Ku antigen.', *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 273(47), pp. 31068–74. doi: 10.1074/jbc.273.47.31068.

Wang, J. L. *et al.* (2018) 'Dissection of DNA double-strand-break repair using novel single-molecule forceps', *Nature Structural & Molecular Biology*, 25(6), pp. 482–487. doi:

10.1038/s41594-018-0065-1.

Ward, J. F. (1994) 'The complexity of DNA damage: relevance to biological consequences.', *International journal of radiation biology*, 66(5), pp. 427–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7983426 (Accessed: 28 August 2019).

Watts, F. Z. and Brissett, N. C. (2010) 'Linking up and interacting with BRCT domains', *DNA Repair*, 9(2), pp. 103–108. doi: 10.1016/j.dnarep.2009.10.010.

Wechsler, T., Newman, S. and West, S. C. (2011) 'Aberrant chromosome morphology in human cells defective for Holliday junction resolution', *Nature*, 471(7340), pp. 642–646. doi: 10.1038/nature09790.

Weterings, E. *et al.* (2009) 'The Ku80 Carboxy Terminus Stimulates Joining and Artemis-Mediated Processing of DNA Ends', *Molecular and Cellular Biology*, 29(5), pp. 1134–1142. doi: 10.1128/MCB.00971-08.

Williams, D. R. *et al.* (2008) 'Cryo-EM Structure of the DNA-Dependent Protein Kinase Catalytic Subunit at Subnanometer Resolution Reveals α Helices and Insight into DNA Binding', *Structure*, 16(3), pp. 468–477. doi: 10.1016/j.str.2007.12.014.

Williams, R. S., Green, R. and Glover, J. N. M. (2001) 'Crystal structure of the BRCT repeat region from the breast cancer-associated protein BRCA1.', *Nature Structural Biology*, 8(10), pp. 838–842. doi: 10.1038/nsb1001-838.

Woodbine, L., Gennery, A. R. and Jeggo, P. A. (2014) 'The clinical impact of deficiency in DNA non-homologous end-joining', *DNA Repair*, 16, pp. 84–96. doi: 10.1016/j.dnarep.2014.02.011.

Wu, D., Topper, L. M. and Wilson, T. E. (2008) 'Recruitment and Dissociation of Nonhomologous End Joining Proteins at a DNA Double-Strand Break in *Saccharomyces cerevisiae*', *Genetics*, 178(3), pp. 1237–1249. doi: 10.1534/genetics.107.083535.

Wu, L. and Hickson, I. D. (2003) 'The Bloom's syndrome helicase suppresses crossing over during homologous recombination', *Nature*, 426(6968), pp. 870–874. doi: 10.1038/nature02253.

Wu, P.-Y. *et al.* (2009) 'Structural and Functional Interaction between the Human DNA Repair Proteins DNA Ligase IV and XRCC4', *Molecular and Cellular Biology*, 29(11), pp. 3163–3172. doi: 10.1128/MCB.01895-08.

Wu, Q. *et al.* (2011) 'Non-homologous end-joining partners in a helical dance: structural studies of XLF–XRCC4 interactions', *Biochemical Society Transactions*, 39(5), pp. 1387–1392. doi: 10.1042/BST0391387.

Wu, Q. *et al.* (2019) 'Understanding the structure and role of DNA-PK in NHEJ: How X-ray diffraction and cryo-EM contribute in complementary ways.', *Progress in biophysics and molecular biology*. doi: 10.1016/j.pbiomolbio.2019.03.007.

Xia, B. *et al.* (2006) 'Control of BRCA2 Cellular and Clinical Functions by a Nuclear Partner, PALB2', *Molecular Cell*, 22(6), pp. 719–729. doi: 10.1016/j.molcel.2006.05.022.

Xing, M. *et al.* (2015) 'Interactome analysis identifies a new paralogue of XRCC4 in non-homologous end joining DNA repair pathway', *Nature Communications*, 6(1), p. 6233. doi: 10.1038/ncomms7233.

Yaneva, M., Kowalewski, T. and Lieber, M. R. (1997) 'Interaction of DNA-dependent protein kinase with DNA and with Ku: biochemical and atomic-force microscopy studies', *The EMBO Journal*, 16(16), pp. 5098–5112. doi: 10.1093/emboj/16.16.5098.

Yang, G. *et al.* (2018) 'Super-resolution imaging identifies PARP1 and the Ku complex acting as DNA double-strand break sensors.', *Nucleic acids research*. Oxford University Press, 46(7), pp. 3446–3457. doi: 10.1093/nar/gky088.

Yin, X. *et al.* (2017a) 'Cryo-EM structure of human DNA-PK holoenzyme.', *Cell research*, 27(11), pp. 1341–1350.

Yin, X. *et al.* (2017b) 'Cryo-EM structure of human DNA-PK holoenzyme', *Cell Research*, 27(11), pp. 1341–1350. doi: 10.1038/cr.2017.110.

Yu, Y. *et al.* (2003) 'DNA-PK phosphorylation sites in XRCC4 are not required for survival after radiation or for V(D)J recombination.', *DNA repair*, 2(11), pp. 1239–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14599745 (Accessed: 29 August 2019).

YU, Y. *et al.* (2008) 'DNA-PK and ATM phosphorylation sites in XLF/Cernunnos are not required for repair of DNA double strand breaks', *DNA Repair*, 7(10), pp. 1680–1692. doi: 10.1016/j.dnarep.2008.06.015.

Zakharyevich, K. *et al.* (2012) 'Delineation of joint molecule resolution pathways in meiosis identifies a crossover-specific resolvase.', *Cell*. NIH Public Access, 149(2), pp. 334–47. doi:

10.1016/j.cell.2012.03.023.

Zelensky, A., Kanaar, R. and Wyman, C. (2014) 'Mediators of Homologous DNA Pairing', *Cold Spring Harbor Perspectives in Biology*, 6(12), pp. a016451–a016451. doi: 10.1101/cshperspect.a016451.

Zhang, F. *et al.* (2009) 'PALB2 Links BRCA1 and BRCA2 in the DNA-Damage Response', *Current Biology*, 19(6), pp. 524–529. doi: 10.1016/j.cub.2009.02.018.

Zhang, Y. (2008) 'I-TASSER server for protein 3D structure prediction', *BMC Bioinformatics*, 9(1), p. 40. doi: 10.1186/1471-2105-9-40.

Zhang, Z. *et al.* (2001) 'The Three-dimensional Structure of the C-terminal DNA-binding Domain of Human Ku70*'. doi: 10.1074/jbc.M105238200.

Zhang, Z. *et al.* (2004) 'Solution Structure of the C-Terminal Domain of Ku80 Suggests Important Sites for Protein-Protein Interactions', *Structure*, 12(3), pp. 495–502. doi: 10.1016/j.str.2004.02.007.

Zhao, H. and Piwnica-Worms, H. (2001) 'ATR-Mediated Checkpoint Pathways Regulate Phosphorylation and Activation of Human Chk1', *Molecular and Cellular Biology*, 21(13), pp. 4129–4139. doi: 10.1128/MCB.21.13.4129-4139.2001.

Zhou, T. *et al.* (2005) 'Deficiency in 3'-phosphoglycolate processing in human cells with a hereditary mutation in tyrosyl-DNA phosphodiesterase (TDP1)', *Nucleic Acids Research*. Narnia, 33(1), pp. 289–297. doi: 10.1093/nar/gki170.

Zhou, Y. *et al.* (2017) 'Regulation of the DNA Damage Response by DNA-PKcs Inhibitory Phosphorylation of ATM', *Molecular Cell*, 65(1), pp. 91–104. doi: 10.1016/j.molcel.2016.11.004.

Zolner, A. E. *et al.* (2011) 'Phosphorylation of polynucleotide kinase/ phosphatase by DNA-dependent protein kinase and ataxia-telangiectasia mutated regulates its association with sites of DNA damage', *Nucleic Acids Research*, 39(21), pp. 9224–9237. doi: 10.1093/nar/gkr647.

Zou, L. and Elledge, S. J. (2003) 'Sensing DNA Damage Through ATRIP Recognition of RPA-ssDNA Complexes', *Science*, 300(5625), pp. 1542–1548. doi: 10.1126/science.1083430.